

SOWING

THE CONSTRUCTION OF HISTORICAL
LONGITUDINAL POPULATION DATABASES



**RADBOUD
UNIVERSITY
PRESS**


EDITED BY
Kees Mandemakers
George Alter
Hélène Vézina
Paul Puschmann

Sowing

Sowing

The Construction of Historical Longitudinal Population Databases

Edited by Kees Mandemakers, George Alter,
Hélène Vézina & Paul Puschmann



Sowing

The Construction of Historical Longitudinal Population Databases

Published by RADBOUD UNIVERSITY PRESS

Postbus 9100, 6500 HA Nijmegen, the Netherlands

www.radbouduniversitypress.nl | radbouduniversitypress@ru.nl

Cover image: *The Sower*, Vincent van Gogh, c. 17-28 June 1888

Cover design: Textcetera

Lay-out: Marja Koster

Print and distribution: Pumbo.nl

ISBN: 978-94-9329-617-6

DOI: 10.54195/BJYF5752

Version: 2023-08

Download book via: www.radbouduniversitypress.nl

Download individual chapters: <https://hlcs.nl/specialissue5>

© 2023, Kees Mandemakers, George Alter, H el ene V ezina and Paul Puschmann

All authors hold copyrights of their own contribution.

**RADBOUD
UNIVERSITY
PRESS**

This is an Open Access book published under the terms of Creative Commons Attribution-Noncommercial-NoDerivatives International license (CC BY-NC-ND 4.0). This license allows reusers to copy and distribute the material in any medium or format in unadapted form only, for non-commercial purposes only, and only so long as attribution is given to the creator, see <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Table of contents

Preface	7
Introduction: Content, Design and Structure of Major Databases with Historical Longitudinal Population Data <i>George Alter, Kees Mandemakers, H��l��ne V��zina</i>	9
I Longitudinal data	
Thank You, Akira Hayami! The Xavier Database of Historical Japan <i>Satomi Kurosu, Miyuki Takahashi, Hao Dong</i>	19
The Demographic Database — History of Technical and Methodological Achievements <i>P��r Vikstr��m, Maria Larsson, Elisabeth Engberg, S��ren Edvinsson</i>	39
The Scanian Economic-Demographic Database (SEDD) <i>Martin Dribe, Luciana Quaranta</i>	53
A Longitudinal Historical Population Database in Asia. The Taiwanese Historical Household Registers Database (1906–1945) <i>Chia-chi Lin, Shu-juo Chen, Ying-chang Chuang, Wen-shan Yang, James Wilkerson, Ying-hui Hsieh, Ko-hua Yap, Yu-lin Huang</i>	69
The 2020 IDS Release of the Antwerp COR*-Database. Evaluation, Development and Transformation of a Pre-Existing Database <i>Sam Jenkinson, Francisco Anguita, Diogo Paiva, Hideko Matsuo, Koen Matthijs</i>	79
The Barcelona Historical Marriage Database and the Baix Llobregat Demographic Database, from algorithms for handwriting recognition to individual-level demographic and socioeconomic data <i>Joana Maria Pujadas-Mora, Al��cia Forn��s, Oriol Ramos Terrades, Josep Llad��s, Jialuo Chen, Miquel Valls-F��gols, Anna Cabr��</i>	101
Slavery in Suriname. A Reconstruction of Life Courses, 1830–1863 <i>Coen W. van Galen, Rick J. Mourits, Matthias Rosenbaum-Feldbr��gge, Maartje A.B., Jasmijn Janssen, Bj��rn Quanjer, Thunnis van Oort, Jan Kok</i>	135
II Family reconstitutions	
PRDH and IMPQ 1800–1849 Quebec Historical Family Reconstitution. Content, Design and Biographical Completeness <i>Lisa Dillon, Marilyn Amorevieta-Gentil, Alain Gagnon, Bertrand Desjardins</i>	159
An Overview of the BALSAC Population Database. Past Developments, Current State and Future Prospects <i>H��l��ne V��zina, Jean-S��bastien Bournival</i>	183

The Ural Population Project. Demography and Culture From Microdata in a European-Asian Border Region <i>Elena Glavatskaya, Julia Borovik, Gunnar Thorvaldsen</i>	199
LINKS. A System for Historical Family Reconstruction in the Netherlands <i>Kees Mandemakers, Gerrit Bloothoof, Fons Laan, Joe Raad, Rick J. Mourits, Richard L. Zijdeman</i>	221
Historical Population Database of Transylvania. Sources, Particularities, Challenges, and Early Findings <i>Luminița Dumănescu, Mihaela Hărăgus, Angela Lumezeanu, Elena Crinela Holom, Nicoleta Hegedűs, Daniela Mârza, Diana Covaci, Ioan Bolovan</i>	259
III Semi-longitudinal data	
The Richness of Italian Historical Demography <i>Marco Breschi, Alessio Fornasin, Matteo Manfredini</i>	279
The Utah Population Database. A Model for Linking Medical and Genealogical Records for Population Health Research <i>Ken R. Smith, Alison Fraser, Diana Lane Reed, Jahn Barlow, Heidi A. Hanson, Jennifer West, Stacey Knight, Navina Forsythe, Geraldine P. Mineau</i>	293
The Development of Microhistorical Databases in Norway. A Historiography <i>Gunnar Thorvaldsen, Lars Holden</i>	313
The Groningen Integral History Cohort Database. Development, Design and Output <i>Richard Paping, Dinos Sevdalakis</i>	335
Geneva. An Urban Sociodemographic Database <i>Michel Oris, Olivier Perroux, Grazyna Ryczkowska, Reto Schumacher, Adrien Remund, Gilbert Ritschard</i>	357
Building an Archival Database for Visualizing Historical Networks. A Case for Pre-Modern Korea <i>Seungmin Paek, Jong Hee Park, Sangkuk Lee</i>	373
The South African Families Database <i>Jeanne Cilliers</i>	389
IV Specific cohorts	
Nominative Linkage of Records of Officials in the China Government Employee Dataset-Qing (CGED-Q) <i>Cameron Campbell, Bijia Chen</i>	407
Building Longitudinal Datasets From Diverse Historical Data in Australia <i>Janet McCalman</i>	435
Reconstructing a Longitudinal Dataset for Tasmania <i>Trudy Cowley, Lucy Frost, Kris Inwood, Rebecca Kippen, Hamish Maxwell-Stewart, Monika Schwarz, John Shepherd, Richard Tuffin, Mark Williams, John Wilson, Paul Wilson</i>	455
Construction of the Finnish Army in World War II Database <i>Ilari Taskinen</i>	483

Preface

Worldwide the number of large historical population databases is on the rise. Whereas just a couple of decades ago most of the databases were from Western countries, an impressive number of databases on non-Western countries has been added to the list. Simultaneously we observe a diversification in the sources that form the basis of those databases. Most of the oldest historical population databases focused initially on classic demographic sources, such as parish registers, vital registration records and population registers, sometimes combined with socio-economic information from, for instance, tax registers and censuses. More recently, we observe also historical population databases based on genealogical data, prison, hospital, burial, government employee and slave registers, as well as other exciting sources. Also, many of the older databases have been enriched by a plethora of new sources. The diversification of sources and the accompanying explosive growth of variables on the life of individuals, their households, families and their living environment is an enormous enrichment for academic research and opens up totally new research lines for historical life course studies. It is especially also a great stimulus for interdisciplinary cooperation of historical demographers with, for example, geneticists, epidemiologists, economists, biologists and medical scientists.

Another exciting trend that we are currently witnessing is an enormous advancement of methods for data entry and record linkage. Thanks to the introduction of, amongst others, Optical Character Recognition (OCR), Handwritten Text Recognition (HTR), Artificial Intelligence (AI), Machine and Deep Learning Techniques, databases can be built and extended ever faster and with ever more precision. In addition, the sheer size of databases has grown massively thanks to advancements in both hard- and software. Last but not least the opportunities for comparative research have grown immensely thanks to the implementation of the Intermediate Data Structure (IDS) (Alter & Mandemakers 2014). In the past, databases used to have all their own structures and unique variable names, complicating comparative analyses on two or more databases. By the introduction of one common database standard for all historical databases, many of the former obstacles to comparative research have been overcome.

For the second conference of the European Society of Historical Demography that was organized in Leuven in 2016, Koen Matthijs, Saskia Hin, Jan Kok & Hideko Matsuo (2016) compiled a collection of essays on the future of historical demography and in the introduction of the volume they called for digging deeper into existing holes of the demographic past, but also for digging new holes. In the period 2020-2023 my colleagues George Alter (ICPSR, University of Michigan), Kees Mandemakers (International Institute of Social History & Erasmus University, Rotterdam) and H  l  ne V  zina (Universit   du Qu  bec    Chicoutimi) have compiled a collection of 23 articles that meticulously describe the history of the construction of large historical population databases, as well as the underlying sources. The collection comprises of descriptions of databases from Europe, North America, Australia, East Asia, South-Africa and the Caribbean. All of these articles have been published in a special issue of *Historical Life Course Studies*, entitled 'Content, Design and Structure of Major Databases with Historical Longitudinal Population Data'.¹

On the occasion of the fifth conference of the European Society of Historical Demography, we have turned the special issue on the construction of large historical population databases in *Historical Life Course Studies* into an edited volume. This volume contains 23 articles representing even more databases. Although all databases contain a variety of sources, coverage and record linkage approaches, a basic distinction can be made into four types: (I) Longitudinal data, (II) Family reconstitutions, (III) Semi-longitudinal data and (IV) Specific cohorts. This distinction was the basis for dividing the book into four sections. Within each section, the databases were further ordered according to the year in which the database took off.

This volume has been published by Radboud University Press, as it provides an exceptional foundation for delving deeper into existing research holes, as well as uncovering new aspects of the demographic past. The collection forms a well-balanced combination of old and established databases that have been used

1 <https://hlcs.nl/specialissue5>

intensively, as well as relatively new and even some brand-new databases on areas, eras, sources and topics that have hardly been explored by historical demographers. We hope that this volume sows many seeds for exciting new research on both old and new historical longitudinal databases.

This edited volume was made possible thanks to generous grants from Radboud University – i.e., the Faculty of Arts, the International Office Arts and the Radboud Group for Historical Demography and Family History, as well as HiDo, the International Network of Historical Demography (Research Foundation Flanders), and the N.W. Posthumus Institute, the Research School for Economic and Social History in the Netherlands and Flanders.

Dr. Paul Puschmann

Assistant Professor of Economic, Social and Demographic History, Radboud Group for Historical Demography and Family History, Radboud University, Nijmegen

Co-editor-in-chief of Historical Life Course Studies, International Institute of Social History, Amsterdam

References

Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for Longitudinal Historical Microdata, version 4. *Historical Life Course Studies*, 1, 1–26. <https://doi.org/10.51964/hlcs9290>

Matthijs, K., Hin, S., Kok, J. & Matsuo, H. (2016). *The Future of Historical Demography. Upside Down and Inside Out*. Leuven & The Hague: Acco.

HISTORICAL LIFE COURSE STUDIES
VOLUME 13 (2023), published 17-07-2023

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

George Alter

University of Michigan

Kees Mandemakers

International Institute of Social History, Amsterdam & Erasmus University Rotterdam

Hélène Vézina

Université du Québec à Chicoutimi

ABSTRACT

In recent years the development of historical databases reconstructing the lives of large populations accelerated. These considerable investments of time and money have greatly expanded possibilities for new research in history, demography, sociology, economics, and other disciplines. This special issue describes the content and design of 23 important historical databases. Authors were given the freedom to discuss a range of practical and technical decisions from evaluating archival sources to crowdsourcing data entry. The most common issue is nominative record linkage, but we find different choices between semi-automatic and fully automatic linkage techniques and various approaches for connecting diverse sources. Some databases describe special problems, like linking Chinese names, handwritten text recognition or the construction of a release in IDS-format. Other databases offer detailed descriptions of sources or discuss prospects for including new datasets.

Keywords: Historical demography, Historical microdata, Life course, Social science history, Record linkage, Standardization historical data, Longitudinal research

DOI article: <https://doi.org/10.51964/hlcs15759>

© 2023, Alter, Mandemakers, Vézina

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 AIMS AND CONTENT

Over the last 60 years several major historical databases with reconstructed life courses of large populations have been launched. The development of these databases is indicative of considerable investments that have greatly expanded the possibilities for new research within the fields of history, demography, sociology, as well as other disciplines. At the annual meeting of the Social Science History Association in Montr al in 2017, the session "Development of Major Databases and their Results from the beginning till now" brought together presentations from some of the largest and most well-established databases with life course data, databases that have also been at the forefront of the development in this field. We were well aware in 2017 of numerous additional databases that had been established around the world in recent decades. In his valedictory speech Kees Mandemakers (2023) made an inventory of a total of 54 databases and even this compilation is not exhaustive.

In order to collect, organize, and then publish information on these major databases in a single collection, invitations were first sent to the leaders of about 25 of these databases. We received in most cases positive and enthusiastic reactions and, when the leaders of a database declined cooperation, it was mostly due to time constraints. We had no specific selection criteria, except that databases had to be actively used and maintained and the primary purpose of the database had to be the (re)construction of individual-level historical life courses. Archived databases, like the Louis Henry dataset (S eguy, 2001), were therefore excluded. Following the first round of invitations, others joined the collective endeavour, expanding the geographic coverage of our collection. We are now very pleased to present contributions representing 24 databases in two special issues of *Historical Life Course Studies*. The number and diversity of databases represented here is truly impressive!

Our overall strategy of describing these major databases resulted in creating two separate special issues. One, *Major Databases with Historical Longitudinal Population Data: Development, Impact and Results*, edited by S oren Edvinsson, Kees Mandemakers and Ken Smith (2023), deals with how the databases contributed to discoveries, responded to changing research questions and facilitated the development of novel lines of inquiry in historical demography and related fields. The present issue focuses on the technical and organizational aspects of these databases, such as their origins and evolution, content and database designs, as well as any setbacks and dependence on external funding. Some recently developed databases appear in this issue with information about both their impact and technical aspects, and several of the impact articles included technical information that is not repeated in this issue. The Chinese database consists of five datasets, three of which are described in the impact issue and two in the technical issue.

This special issue presents 23 articles, including seven databases with counterparts in the impact issue. The six databases appearing in the impact one are the Historical Chinese Micro Database, the Demographic Database of Ume , the Utah Population Database, the Scanian Economic Demographic Database of Lund (SEDD), the Norwegian Historical Micro Database and the Antwerp COR*-database. The Historical Sample of the Netherlands (HSN), covered in the first published article of the impact issue, does not have a technical counterpart, but basic information about the HSN can be found in that article and the article in this special issue on the LINKS database, which is an offshoot of the HSN.

The 24 databases described in these issues represent a larger number of datasets. Whether the database is described as one or many datasets depends primarily on the strategy of the database managers. In general, when data from multiple sources are linked and integrated, the database is considered one big dataset. This is usually the result of a strategy that extended a core database with other data. This could be the result of systematic planning like the Ume  database, a result of the possibilities of crowd sourcing like the Tasmanian database, the result of funding by researchers to extend the dataset with specific data like the HSN, or a combination of these approaches.

The technical issue allows authors freedom to discuss a wide range of issues. The most common issue is nominative record linkage, but we find different choices between semi-automatic and fully automatic linkage techniques (LINKS, BALSAC, SEDD) and different approaches for connecting diverse sources (Utah, Tasmania). Some contributions describe special problems, like linking Chinese names, handwritten text recognition (Barcelona), and the construction of a release in IDS-format. Other contributions offer detailed descriptions of sources (Taiwan, Korea, Finland, Suriname) or discuss prospects for including new datasets.

2 EARLIER INITIATIVES

This is by far the most comprehensive but not the only collection of technical descriptions of databases with micro-data on historical populations. Twenty years ago, an overview of historical databases, the *Handbook of International Historical Microdata for Population Research* (Kelly Hall, McCaa, & Thorvaldsen, 2000), was published by the Minnesota Population Center. It included 16 databases of which five are included in this special issue. These are the databases from Norway, Sweden (Umeå), the Netherlands (HSN), Italy and Canada (PRDH). Most of the databases are not included here, because they concentrate on relatively modern 20th century census data. Four databases with historical census data were not included because they have not linked their data into life courses (like the UK dataset and Stockholm dataset), had no time to participate (Denmark), or are already described in an extensive way, like the IPUMS database initiated by Steven Ruggles (Helgertz et al., 2022; Roberts et al., 2003; Ruggles, 2014; Sobek et al., 2011).

The *Handbook* was a product of IMAG, the International Microdata Access Group, that was formed to realize international collaboration between researchers working with historical micro-data. The group was formed at the 1998 Social Science History Association conference by a group of researchers who desired discussions focused on the problems and potential of micro-data. The IMAG group concentrated on census data gathered by the IPUMS group, and the first IMAG workshop was hosted by the University of Ottawa in 1999 (Dillon, 2000). In November 2003, a second IMAG workshop in Montréal went a step further by concentrating on record linkage, i.e., the ways multiple appearances of the same persons and households were linked in various databases (Dillon & Roberts, 2006).

Since these first initiatives, cooperation between historical micro-databases intensified enormously. In May 2001, the International Institute of Social History organized a workshop on the results and practices of large databases, resulting in an overview of best practices for these databases (Mandemakers & Dillon, 2004). This was followed by a full day session on "New sources for historical demographic research" with four panels, organized by the International Commission for Historical Demography at the World History Conference in Sydney in 2005.

In March 2006 a second workshop, "Disseminating and Analyzing Longitudinal Historical Data", took place at IISH Amsterdam. Although participants recognized the complex nature of longitudinal databases, the workshop ended with a consensus on how to make progress. First, it was agreed that standardization in the products of the different databases would help researchers enormously. Second, an intermediate data structure (IDS) was proposed to mediate between the original databases and the data sets required for analysis. On May 2008, the Inter-university Consortium for Political and Social Research (ICPSR) hosted a planning group to continue working on the IDS. This resulted in a model for data sharing, which was presented to an open meeting of historical databases at the Social Science History Association meeting in Miami, October 2008 (Alter & Mandemakers, 2014; Alter, Mandemakers, & Gutmann, 2009). Part of the 2006 workshop was an initiative to publish questionnaires with key information about the databases the participants were representing. This included the Historical Database of the Liège Region (Belgium), Scania Database (Sweden), Registre de la population du Québec ancien (PRDH), Historical Sample of Flanders, Demographic Database Umeå (Sweden), Victorian Scotland database, Connecticut River Valley Project (USA), Texas Longitudinal Data Project (USA), Migration Database (USA, based on genealogies), Danjuro Database Japan, Historical Sample of the Netherlands (HSN), Koori Health Research Database (KHRD) 1855–1930 (Australia), Melbourne Lying-In Hospital Cohort: 1857–1900, Utah Population Database, Geneva Database, IPUMS database (census USA), Norwegian Census Database.

The initial IDS working group was succeeded by the ESF funded European Historical Population Sample Network project (EHPS-Net), which ran from 2011 to 2016. This project gathered almost all historical micro-databases with a European background. Networking activities around historical population databases and the IDS continued with the LONGPOP project, a Marie Curie Innovative Training Networks project from 2017 to 2020 on "Methodologies and Data mining techniques for the analysis of Big Data based on Longitudinal Population and Epidemiological Registers".

One of the spin-offs of the EHPS-network is the journal *Historical Life Course Studies* which is publishing these two special issues on the technical aspects and impacts of large historical population databases. Both are the result of the already mentioned session on the "Development of Major Databases and Their Results from the Beginning till Now" at the Social Science History Conference in Montréal 2017,

which had presentations from the Utah Population Database (UPDB), the Historical Sample of the Netherlands (HSN), the Demographic Database of Ume  (DDB) and the BALSAC database of the Universit  du Qu bec   Chicoutimi.

The database questionnaires collected for the IISH conference in 2006 were taken over by the EHPS-network. The number of participating database increased to 32 (<https://ehps-net.eu/databases>), and much more detail was added. This special issue offers a new overview by providing an opportunity to update and expand information on many of the databases previously described and by including presentations on recent initiatives. Fourteen of these 29 databases are included in the technical or impact special issue. Four databases were not included because their authors could not produce an article (TRA database France, Denmark, Li ge & Verviers region, Isle of Skye which became a subset of Digitizing Scotland). Census data projects that have not been linked into life courses, like MOSAIC and the Canada Family project, are also omitted. Eight databases are missing, because contact with the database managers has been lost or the database was judged too marginal for inclusion in these special issues (see Mandemakers (2023) for estimates of the numbers of persons and households/families included). However, we include a number of databases that did not appear in the questionnaires. Databases developed since 2006 describe Chinese Officials, the military in Finland, slavery in Suriname, family reconstitutions in the Netherlands, Barcelona, Taiwan, Korea, South-Africa, and the Urals. We also have articles on older datasets not described in the questionnaires, like databases on Utah, Tasmania, China, and the Xavier collection in Japan.

3 TYPOLOGY OF INCLUDED DATABASES

Life course databases may be divided into three types: (I) longitudinal data, (II) family reconstitutions, and (III) semi-longitudinal data. The first type, 'pure' longitudinal data, are based on sources like population registers that record continuous observation of vital events (births, marriages, and deaths) as well as migration, such as HSN, SEDD Lund, DDB Ume , and the Antwerp COR*-database. In contrast, family reconstitutions, such as LINKS, must be analyzed under the restrictive rules developed by Louis Henry, because they only include vital events (Fleury & Henry, 1956, 1985; Henry, 1970; Henry & Blum, 1988). Semi-longitudinal databases (e.g., Utah, Norway and China) combine vital registration with censuses, taxes, and other nominative lists that identify the population under observation (see Alter, 2019).

We can also divide databases by geographic coverage (Mandemakers, 2023). Most historical databases cover only selected communities or specific areas within a country (e.g., Barcelona, Geneva, Qu bec, Tasmania, Utah). Several databases are national in scope, such as the HSN, LINKS, and the Norwegian database, but national databases may focus on specific cohorts, like government officials in China, babies born in a charity hospital, aboriginals or deported convicts (Australia), and Finnish and Australian military recruits.

The following table presents an overview of all 32 databases/datasets in our two special issues. We see that five of the included databases are typical of family reconstitution, and nine are based on fully longitudinal sources. We classify a majority of the databases as "semi-longitudinal", because they use censuses or other nominative lists, often in combination with vital registration. Within this group of semi-longitudinal datasets we distinguish a subgroup of five datasets having only data from linked censuses. Nationwide coverage is available in five databases, 19 databases cover only part of a country, and eight are nationwide but only include a special cohort.

Table 1 *Overview of all included databases*

Nature_basic	Coverage	Name	Country	Technical	Impact
Longitudinal	Nationwide	Historical Sample of the Netherlands	Netherlands		X
Longitudinal	Nationwide	Historical Databae of Suriname	Suriname	X	
Longitudinal	Regional	Antwerp COR*-database	Belgium	X	X
Longitudinal	Regional	Baix Llobregat Demographic Database (BALL)	Spain	X	
Longitudinal	Regional	POPLINK DDB Umeå	Sweden	X	X
Longitudinal	Regional	POPUM DDB Umeå	Sweden	X	X
Longitudinal	Regional	Scanian Economic Demographic Database	Sweden	X	X
Longitudinal	Regional	Taiwan Historical Household Registers Database	Taiwan	X	
Longitudinal	Regional	Xavier Database of Japan	Japan	X	
Family Reconstitution	Nationwide	LINKS	Netherlands	X	X
Family Reconstitution	Regional	Registre de la Population du Québec Ancien (RPQA)	Canada	X	
Family Reconstitution	Regional	Historical Population Database of Transylvania	Rumania	X	
Family Reconstitution	Regional	Ural Population Project	Russia	X	
Family Reconstitution	Regional	BALSAC Population Database	Canada	X	
Semi-longitudinal	Nationwide	South African Families Database	South-Africa	X	
Semi-longitudinal	Regional	Utah Population Database	USA	X	X
Semi-longitudinal	Regional	Barcelona Historical Marriage Database	Spain	X	
Semi-longitudinal	Regional	Integral History Project Groningen	Netherlands	X	
Semi-longitudinal	Regional	Italian Historical Population Database	Italy	X	
Semi-longitudinal	Special cohort	China Government Employee Datasets-Qing (CGED-Q) Jinshenlu (JSL) and Examination Records (ER)	China	X	
Semi-longitudinal	Special cohort	China Multigenerational Panel Database-Imperial Lineage	China		X
Semi-longitudinal	Special cohort	Diggers to Veterans	Australia	X	
Semi-longitudinal	Special cohort	Finnish Army in World War II Database	Finland	X	
Semi-longitudinal	Special cohort	Founders and Survivors (Linked datasets)	Australia	X	
Semi-longitudinal	Special cohort	Founders and Survivors (Ships cohort)	Australia	X	
Semi-longitudinal	Special cohort	Koori Health Database	Australia	X	
Semi-longitudinal	Special cohort	Melbourne Lying-In Hospital Cohort	Australia	X	
Semi-longitudinal Census	Nationwide	Norwegian Historical Population Register, 1800–1964	Norway	X	X
Semi-longitudinal Census	Regional	China Multigenerational Panel Database-Liaoning	China		X
Semi-longitudinal Census	Regional	China Multigenerational Panel Database-Shuangcheng	China		X
Semi-longitudinal Census	Regional	Geneva Demographic Database	Switzerland	X	
Semi-longitudinal Census	Regional	Korean Historical Archives Visualization Network Database	Korea	X	

Each database represented in this issue has its own unique genesis that is well described in the various papers. For example, the launch of the DDB at Ume a University was initially motivated by an interest in the development of literacy. For the Utah Population Database, the impetus was the value of genealogies and family histories for genetics and medicine. At the same time, several common elements and developmental arcs connect these distinct databases. In many respects, the scientific relevance for the development of historical population databases rests on the shoulders of giants who championed quantitative history and the history of the ordinary person. This includes members of the Cambridge Group for the History of Population and Social Structure (Wrigley, Davies, Oeppen, & Schofield, 1997), the Annales School with its advocacy of social history (S eguy, 2016), and the proponents of the life course perspective arguing for the plasticity of human development and the role of history (Kok, 2007). With these intellectual foundations as bedrock, technological advances proved to be a catalyst for accelerating the insights of quantitative history by digitizing archival records and through record linking methodologies that reveal the diversity of human life courses.

The origins of historical longitudinal databases vary, but six in our issue originate from the 1970's when computers and software first facilitated data entry, processing and database management. These are the Xavier database of Japan, the Demographic Database Ume a, the Utah Population Database and the Norwegian Historical Data Centre in Troms o, as well as the two databases about Qu ebec (the Registre de la population du Qu ebec ancien (Universit e de Montr eal) and the BALSAC database (Universit e du Qu ebec   Chicoutimi)). These projects created a legacy through numerous publications, large numbers of trainees, and the development of stable and reliable infrastructures. In the following two decades, they were followed by new databases like the SEDD database in Sweden, the Chinese datasets, and the HSN database in the Netherlands. Although databases continue to be added in western Europe (e.g., LINKS, Antwerp COR*-database, and Barcelona), the most impressive expansion has been in other parts of the world, including Asia, Australia, South Africa, and Eastern Europe. The technical special issue describes new challenges encountered in these areas as well as the opportunities offered by new technologies, like machine learning and natural language processing. While the expansion of these infrastructures is impressive and benefits the research community broadly, significant portions of the globe are not yet represented, largely due to lack of resources needed to create and maintain complex databases.

REFERENCES

- Alter, G. (2019). The evolution of models in historical demography. *Journal of Interdisciplinary History*, 50(3), 325–362. doi: [10.1162/jinh_a_01445](https://doi.org/10.1162/jinh_a_01445)
- Alter, G., Mandemakers, K., & Gutmann, M. (2009). Defining and distributing longitudinal historical data in a general way through an intermediate structure. *Historical Social Research*, 34(3), 78–114. doi: [10.12759/hsr.34.2009.3.78-114](https://doi.org/10.12759/hsr.34.2009.3.78-114)
- Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies*, 1, 1–26. doi: [10.51964/hlcs9290](https://doi.org/10.51964/hlcs9290)
- Edvinsson, S., Mandemakers, K., & Smith, K. R. (2023). Introduction: Major databases with historical longitudinal population data: Development, impact and results. *Historical Life Course Studies*, 13, 186–190. doi: [10.51964/hlcs14840](https://doi.org/10.51964/hlcs14840)
- Dillon, L. Y. (2000). Preface. The origins of IMAG: The International Microdata Access Group. In P. Kelly Hall, R. McCaa & G. Thorvaldsen (Eds.), *Handbook of international historical microdata for population research* (XI–XIV). Minneapolis: Minnesota Population Center. Retrieved from https://international.ipums.org/international/microdata_handbook.shtml
- Dillon, L. Y., & Roberts, R. (2006). Introduction: Longitudinal and cross-sectional historical data: Intersections and opportunities. *History and Computing*, 14(1–2), 1–7.
- Flcury, M., & Henry, L. (1956). *Des registres paroissiaux   l'histoire de la population: manuel de d epouillement et d'exploitation de l' tat civil ancien* [From parish registers to the history of the population: Manual for counting and exploitation of the ancient civil status]. Paris: Editions de de l'Institut National d'Etudes D emographiques.
- Flcury, M., & Henry, L. (1985). *Nouveau manuel de d epouillement et d'exploitation de l' tat civil ancien* [New manual for counting and using old civil status] (3rd ed.). Paris: Institut National d'Etudes D emographiques.

- Kelly Hall, P., McCaa, R. & Thorvaldsen, G. (Eds.). (2000). *Handbook of International Historical Microdata for Population Research*. Minneapolis: Minnesota Population Center. Retrieved from https://international.ipums.org/international/microdata_handbook.shtml
- Helgertz, J., Price, J., Wellington, J., Thompson, K. J., Ruggles, S., & Fitch, C. A. (2022). A new strategy for linking U.S. historical censuses: A case study for the IPUMS multigenerational longitudinal panel. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 55(1), 12–29. doi: [10.1080/01615440.2021.1985027](https://doi.org/10.1080/01615440.2021.1985027)
- Henry, L. (1970). *Manuel de démographie historique* [Handbook of historical demography]. Paris: Droz.
- Henry, L., & Blum, A. (1988). *Techniques d'analyse en démographie historique* [Analytical techniques in historical demography] (2nd ed.). Paris: Institut National d'Etudes Démographiques.
- Kok, J. (2007). Principles and prospects of the life course paradigm. *Annales de démographie historique*, 113(1), 203–230. doi: [10.3917/adh.113.0203](https://doi.org/10.3917/adh.113.0203)
- Mandemakers, K. (2023). “You really got me”. *Ontwikkeling en toekomst van historische databestanden met microdata* [Development and future of historical databases with microdata] (Valedictory speech). Rotterdam: Erasmus University Rotterdam. doi: [10.25397/eur.23256467](https://doi.org/10.25397/eur.23256467)
- Mandemakers, K. & Dillon, L. (2004). Best practices with large databases on historical populations. *Historical Methods*, 37(1), 34–38. doi: [10.3200/HMTS.37.1.34-38](https://doi.org/10.3200/HMTS.37.1.34-38)
- Roberts, E., Ruggles, S., Dillon, L. Y., Gardarsdóttir, Ó., Oldervoll, J., Thorvaldsen, G., & Woollard, M. (2003). The North Atlantic Population Project. An overview. *Historical Methods*, 36(2), 80–88. doi: [10.1080/01615440309601217](https://doi.org/10.1080/01615440309601217)
- Ruggles, S. (2014). Big microdata for population research. *Demography*, 51(1), 287–297. doi: [10.1007/s13524-013-0240-2](https://doi.org/10.1007/s13524-013-0240-2)
- Séguy, I. (2001). *La population de la France de 1670 à 1829: l'Enquête Louis Henry et ses données* [The population of France from 1670 to 1829: The Louis Henry survey and its data]. Paris: INED.
- Séguy, I. (2016). The French school of historical demography (1950–2000). In: A. Fauve-Chamoux, I. Bolovan, & S. Sogner (Eds.). *A global history of historical demography. Half a century of interdisciplinarity* (pp. 257–276). Bern: Peter Lang.
- Sobek, M., Cleveland, L., Flood, S., Kelly Hall, P., 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. doi: [10.1080/01615440.2011.564572](https://doi.org/10.1080/01615440.2011.564572)
- 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. doi: [10.1017/CBO9780511660344](https://doi.org/10.1017/CBO9780511660344)

I Longitudinal Data



HISTORICAL LIFE COURSE STUDIES
VOLUME 11 (2021), published 11-11-2021

Thank You, Akira Hayami!

The Xavier Database of Historical Japan

Satomi Kurosu

Reitaku University

Miyuki Takahashi

Rissho University

Hao Dong

Peking University

ABSTRACT

This article introduces the Xavier database, one of the major sources for studying historical populations in Japan. The database consists of 162 years of annual observations for 28,105 individuals living in three villages and one town of the current Fukushima prefecture between 1708 and 1870. We review the extensive efforts of the founder of Japanese historical demography, Akira Hayami, and his group in collecting, transcribing, coding, and finally making local population registers into this database for demographic analysis. We discuss the studies that flourished domestically and internationally using the data in the last two decades, followed by the discussion of current and promising development.

Keywords: Japan, Historical demography, Household registers, Longitudinal data

DOI article: <https://doi.org/10.51964/hlcs11113>

© 2021, Kurosu, Takahashi, Dong

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

"Thank you, Francisco Xavier!" wrote Akira Hayami (1928–2019), the founder of Japanese historical demography in 1979. Without St. Francisco de Xavier who first brought Christianity to Japan in 1549, the *Shumon-Aratame-Cho* (SAC; religious affiliation investigation registers) would never have started. SAC was the "chance by-product of the Tokugawa *Bakufu's* fear and loathing of Christianity" (Hayami, 1979) which serves as the main source for Japanese historical demography. This article introduces the source, the construction, as well as the impact of the *first* digitized data set in Japanese historical demography. The data set was named after Xavier who was one of the founders of the Society of Jesus.

The construction of individual-level longitudinal data based on household registers, SAC and similar *Ninbetsu-Aratame-Cho* (NAC) for early modern Japan has opened up arrays of possibilities for investigating the demographic behavior of commoners in the Tokugawa period (1603–1867). Japanese historical demography has come a long way since Akira Hayami's application of the method of "family reconstitution" to Japanese household registers in the late 1960s. Hayami's lifetime collection of materials on historical demography from his earlier offices including Keio University, International Research Center for Japanese Studies (Nichibunken), and Reitaku Tokyo Center, are now hosted as the "Reitaku Archives" at Reitaku University, organized and maintained by the Population and Family History Project (PFHP) headed by Satomi Kurosu.

This article first introduces the sources used for the Xavier Database and the studies of Japanese historical demography in general. Particular attention is paid to how Hayami's collection of historical records, those which he called "treasure of humankind", were transcribed and put into a format with linked annual individual/household information. Next, we discuss the process of coding and construction of the Xavier Database, the first systematic database Hayami initiated in late 1980s. While there are numerous datasets and sources available across Japan by now, the sources for northeastern communities used for the Xavier Database are known to be the most detailed. We review the Xavier data, focusing on three villages (Niita, Shimomoriya, Hidenoyama) and one town (Koriyama) in the current Fukushima prefecture. Finally, we discuss how the studies based on the dataset advanced our understanding of people's lives through the analysis of behavior and organization of individuals, married couples, and households, as well as their influence on the studies of historical demography and family history.

2 SOURCES

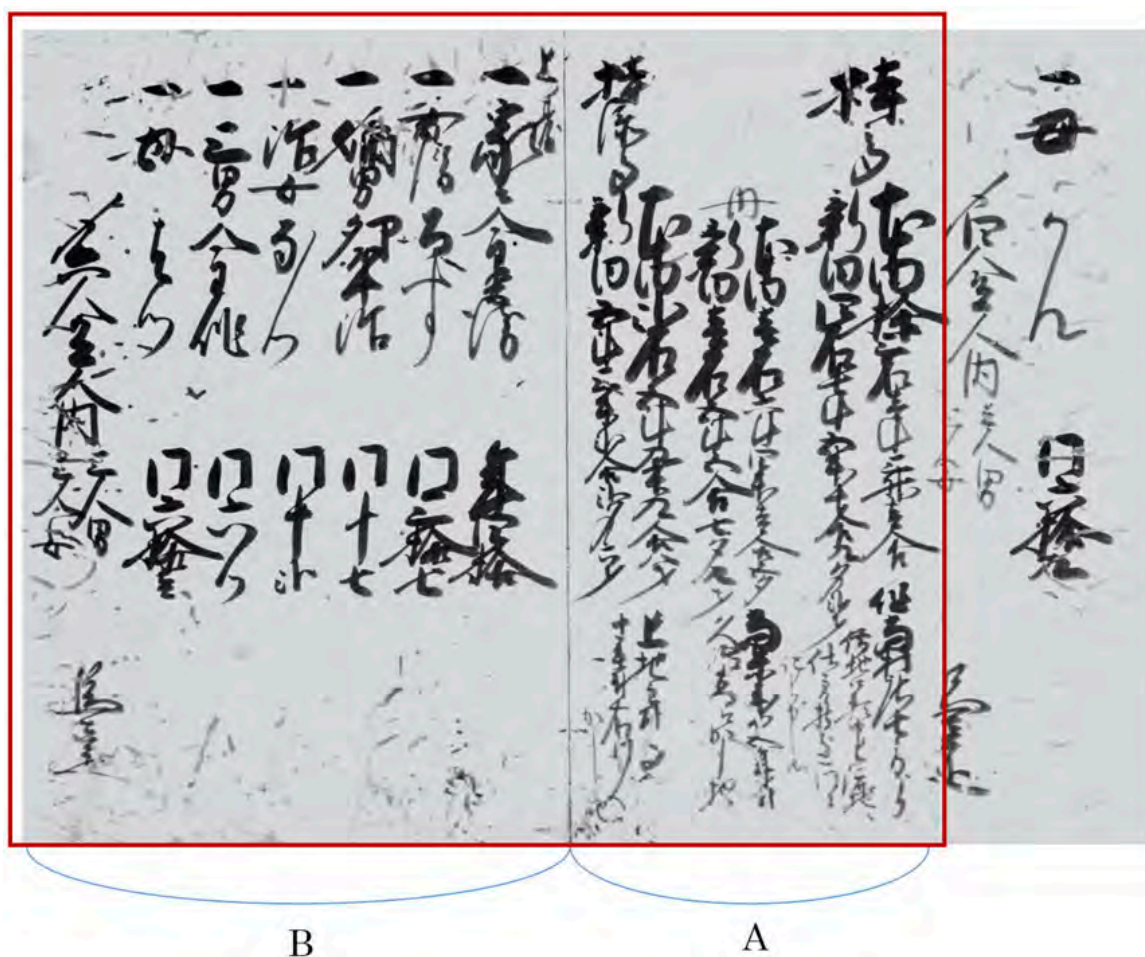
Two primary sources used for the Xavier Database as well as studies of historical demography in Japan are SAC and NAC.¹ Both of them are household registers and include basic information such as an individual's name, age, and relationship to the household head. A household can include not only kin members but also non-kin and servants. Notably, households can be registered even without any resident (e.g., actual residents left for work migration).

SAC was a religious investigation initiated around 1638 by the Tokugawa government as a measure to prevent the entry and spread of Christianity (Cornell & Hayami, 1986; Hayami, 1979). The quality, dates of compilation, and availability of SAC vary depending on the village and region. Some SACs are based on *de jure* information and contain excess numbers of elderly persons, who, for example, out-migrated and possibly died elsewhere but are still listed in their households of origin. NAC preceded SAC and was a type of population register in nature similar to the contents of SAC, except that NAC excludes information on religious affiliation of individual villagers (Hayami, 2001). Meanwhile, NAC tends to record detailed information on population *de facto*. However, in most cases, it is not easy nor practical to distinguish between the two because the two sources are often indistinguishable and are indiscriminately compiled resulting in the title of the documents *Shumon-Ninbetsu-Aratame-Cho*.

1 Other sources not discussed in this article include temple death registers (*Kako-Cho*), pregnancy records (*Kainin-Kakiage-Cho*), records of population increase and decrease (*Zogen-Cho*), and *Hokonin-Uke-Jo*, a register of servants (*hokonin*) that recorded the contract detailing the type of service and length. Availability and accessibility of these sources vary greatly by region. When *Zogen-Cho* and *Hokonin-Uke-Jo* were available, they were matched with SAC/NAC to either check the content or to supply additional information for analysis.

It should be noted that unlike records of elites in other societies, both SAC and NAC documented "commoners". They include peasants, fishermen, merchants, etc., who were the majority of the population in 17th- to early 19th-century Japan. Tokugawa period was a highly stratified society. The elite bureaucrats and administrators during the Tokugawa period (*samurai* class), as well as the members of the Imperial court whose residency was segregated, were recorded separately from the commoners.² The documents available today are copies of the SAC/NAC, kept in the hands of village officials after submitting them to local lords. The officials kept the register copies in order to add annotations for changes of individual vital events (e.g., birth, death, marriage, service) that would occur until the next survey. These annotations provide valuable information of vital records. However, the level of details in such registers differs by local government practices. For some *han* (domains, administrative units governed by *daimyo*, territorial lords), SAC was not always done every year. For other domains, only those after certain ages were registered (e.g., after age 15 for Maeda domain, after age 8 for Kishu and Hiroshima domains). The more detailed listings include origins and destinations of migrants with reasons (e.g., marriage, adoption, service), as well as household landholdings. According to Hayami (2001, p. 25), some of the best sources, in terms of quality and length (continuing more than one century with very few years missing in-between) come from the villages in Nihonmatsu domain (in current Fukushima prefecture), and this is where the very first attempt to systematically construct a database was made.

Figure 1 Original "Ninbetsu-Aratame-Cho": Household No.112 in Niita, year 1763



Source: Microfilm, Reitaku Archives, Population and Family History Project, Reitaku University.

Notes: NAC recorded information of all households in a village/town sequentially in one volume (book bound with Japanese paper) per year. The red square indicates one unit of household. Traditionally, the page reads from right to left. "A" lists household landholdings and lands being leased/rent. "B" lists each member of household. See text for details.

2 For studies of *samurai* demography and family, genealogies are often used.

Figure 1 shows two pages of 1763 NAC in the village of Niita. The red square indicates one household unit. Reading from right to left, the unit starts with "A" which lists household landholding as well as lands being leased/rented, where "B" lists each household member with information about the relationship to the household head, name, and age. In this example, six household members are included: the household head (age 40), his wife (37), his first child-son (17), his second child-daughter (12), his third child-son (6), and his mother (61). The indication of the birth order of children suggests the importance of sibling hierarchy in the area. The column on the left side sums number of males and females in this household followed by number of horses.

3 COLLECTION AND TRANSCRIPTION OF REGISTERS

The collection of SAC/NAC was initiated and led by Akira Hayami in the late 1960–80s. It involved locating/finding SAC/NAC, contacting the holder, getting permission, and microfilming the materials. A great number of people of Hayami's research group have been involved in the collection of sources for over four decades. The sources for the Nihonmatsu domain in the Xavier Database were collected in local archives and private homes. The microfilms were then printed and transcribed into forms and numbers.

Since SAC/NAC registers provide annual information of household and individual life courses spanning up to 200 years, it was not easy to organize. Hayami's first trial was a manual organization of cards that tracked individual households per year (*kohyo*, not shown). Later on, he came up with the Basic Data Sheet (BDS) method of transcribing and organizing information of households for 25 years per sheet (see Figure 2). Hayami recalled that it was an "innovation" as it finally allowed researchers to track what was happening to the individuals and households "longitudinally". And, indeed, BDS was a breakthrough as a method for compiling detailed longitudinal information. It also opened up the opportunity for anyone with knowledge of contemporary Japanese to understand the content of these household registers, which otherwise would be hard to decipher without extensive training in early modern calligraphy. The annual information of a household member (name, relationship, age) is transcribed on the left panel of a BDS. Any movements or changes of status, including birth, death, marriage/adoption, and name changes, are annotated with a symbol and described at the bottom of the BDS. Any household members currently not residing in the household are placed in the right panel of the BDS with information of their locations and the reasons for not residing (for example, working as servant at other village/household). At the very right column of a BDS, the number of horses and the landholdings of the household are entered. SAC/NAC of the same household in the consecutive years are matched and transcribed following the previous year. Organized in this way, both the information of individuals and households can be tracked for the entire observation period.

The linkage of households from one year to the next was done by matching the order of appearance of the households and checking names and ages of household members. Once the BDS was written, individuals were given unique identifiers. When the linking of individual information was done manually this was done by coders; for example, matching and giving the same ID to an individual who left the household of origin and entered another household via marriage/service. Individuals were matched as long as the move took place within the village. Thus, BDS became a valuable source of longitudinal demographic information that follows intriguing life histories of commoners — like reading a biography.

Because it was before the computer revolution, most of Hayami's initial work was done manually extracting information from BDS to construct individual life courses and family. The ITS (individual tracing sheet, see Figure 3) was Hayami's invention for tracking individual movements. And, the FRF (Figure 4) follows the famous "family reconstitution form" of the French historical demographer, Louis Henry. These forms are still kept at the Reitaku Archives, although they are taken over by computer calculation and are no longer used. While Hayami has been transcribing and organizing materials that he and his group collected all over Japan, his research concentrated in the areas of Suwa (Hayami, 1973) and Nobi (Hayami, 1992) — central Japan. He rarely used BDS of the Nihonmatsu domain. Instead, Saeko Narimatsu single-handedly transcribed the original NACs to BDS. With the help of research assistants, she linked individual information from SAC/NAC to BDS and further cross-checked qualitative information of the area from local documents. Based on this laborious work, she published two books on Shimomoriya (Narimatsu, 1985) and Niita (Narimatsu, 1992), which are must-reads for later researchers studying the area.

Figure 2 BDS (Basic Data Sheet): Household No.112 in Niita, years 1751–1775

The table is a grid with columns numbered 1 to 25. The rows represent different years and household members. Key sections include:

- Header:** 上代 (Previous Generation), 名 (Name), 姓 (Surname), 戸の主親 (Head of Household), 性別 (Sex), 戸の主親 (Head of Household).
- Year Column:** 寛延 4年 (1751), 宝曆 2 (1752), 安永 2 (1753), 1754, 1755, 1756, 1757, 1758, 1759, 1760, 1761, 1762, 1763, 1764, 1765, 1766, 1767, 1768, 1769, 1770, 1771, 1772, 1773, 1774, 1775.
- Relationships:** 父 (Father), 母 (Mother), 妻 (Wife), 子 (Child), 孫 (Grandchild), 兄弟 (Brother/Sister), 嫁 (Daughter-in-law), 婿 (Son-in-law), 孫 (Grandchild).
- Migration information:** 遷 (Migration), 入 (Entry), 出 (Exit), 亡 (Death).
- Details of land lease/rent:** 田 (Rice field), 畑 (Field), 園 (Garden), 池 (Pond), 山 (Mountain), 水 (Water), 石 (Stone), 土 (Soil).
- Events/change of name:** 名 (Name), 姓 (Surname), 字 (Courtesy name), 号 (Style name).

Source: Copy of BDS, Reitaku Archives, Population and Family History Project, Reitaku University.

Note: The blue arrow on the left refers to the household of NAC shown in Figure 1, an example of the NAC source. See text for details.

Map 1 shows the original data collection and ongoing database construction efforts spreading nationwide in Japan, as well as the area of the Xavier Database introduced in this article. The steps described above are also shown in Table 1 in relation to the figures and maps. The Xavier Database is one of the first series of Japanese historical demography datasets constructed by Akira Hayami and the members of the Eurasia Project Japan (EAP-J 1995–2000)³ and is the most detailed and vigorously used in the last decades.⁴ Details of Xavier construction are discussed in the next section.

Meanwhile, there are numerous village records entered during and after the time of EAP-J for both longitudinal (red dots in Map 1) and cross-sectional data (orange dots in Map 1) types used in historical demography and family studies. Also, there are plenty of understudied BDS (green dots in Map 1) and microfilms (blue dots in Map 1), now hosted at the PFHP. Since longitudinal data from various locations were not in the same format as the data of the Xavier Database, they were not integrated into one database during the time of EAP-J. Cross-sectional data include those with only one- or two-year SAC records. While those cross-sectional data have not been utilized for longitudinal research in Japanese historical demography, it has a potential for investigating regional variations as well as fertility dynamics (e.g., Drixler, 2013; Kurosu, 2008).

3 Eurasia Project of Population and Family History was funded by the Japanese Ministry of Education Grant-in-Aid for Creative Basic Research (PI Akira Hayami). Official members in Japan included at least 37 scholars from various disciplines (history, sociology, anthropology, and information science). Further collections of original sources and numerous data entry, data base construction, as well as collaborative studies were pursued during this period and after. The international collaboration of Eurasia Project (EAP) continued until recently (three volumes are found here: <https://mitpress.mit.edu/books/series/eurasian-population-and-family-history>).

4 Another useful database is constructed by a member of the Japanese Eurasia Project, Hiroshi Kawaguchi, DANJURO (<http://www.danjuro.jp/>). The database includes Aizu villages in Fukushima and is publicly released (registration required).

There are numerous studies using BDS sheets of areas other than the northeastern ones included in the Xavier Database. To name a few, Hayami's own work of Suwa (1973) and Nishijo and the surrounding villages (1992) and Cornell's work on Yokouchi (1981) in central Japan; and a recent work by Nakajima utilizing a fishing village of Nomo in southwest Japan (2016).

Figure 3 ITS (Individual Tracing Sheet): A female born in 1722 until her death in 1790, Shimomoriya

地域コード: 下守屋		PR No. 78-006		出現理由: B	中間消滅回数: 2	最終消滅理由: D	性別: F		
年代: 1722		年代: 1790		出身: 下守屋		宗派:			
分類I: ●	分類II:	出身: 下守屋		宗派:					
年代	年齢	理由	続柄	備考	年代	年齢	理由	続柄	備考
N 名前 (名前)					N 名前 (名前)				
1.	1721-	1	B 女子	レ け	2.	-			
2.	-				3.	-			
3.	-				4.	-			
F 所属する家族 (奉公・出稼の場合を除く)									
1.	1721-	1	B 女子	HC No 78	2.	1747	27	V 妻	HC No 78A
2.	-			HC No	3.	-			HC No
3.	-			HC No	4.	-			HC No
P 家族内の位置 (奉公・出稼の場合を除く)									
1.	1721-	1	B 女子	HC No 78	2.	1747-	27	V 妻	HC No 78A
2.	1774-	54	親祖母	HC No 78A	3.	-			HC No
3.	-			HC No	4.	-			HC No
4.	-			HC No	5.	-			HC No
B 出生 (父のPRNo) (母のPRNo) (FRFNo) D 死亡									
1.	1721-	7	78-001	78-002	78-01	1790-	70		母
2.	男・女	間隔	年	出生時年齢: 父= 29	2.	同一年次の家族の死亡:			
3.	私生:	双生:		母= 26	3.	前後の年の家族の死亡:			
M 結婚 (FRF) (相手のPR又は出身) DV 離婚 (死別を含む) (離婚後の行動)									
1.	1731-	11	78-03	78-008	1.	1787-	67	MD	再婚セリ
2.	-				2.	-			
3.	-				3.	-			
4.	-				4.	-			
R 養子 (結婚を伴わない場合) RC 養子の不嫁戻り									
1.	-				1.	-			
2.	-				2.	-			
E 奉公・出稼 (種別) (出稼先) 変 EC 奉公・出稼戻り (出稼先)									
1.	1760-	40	E出	78 → 郡山	2.	-			
2.	1761-	41	E出	郡山 → 大槻	2.	-			
3.	1762-	42	E出	大槻 → 富岡	2.	1764-	44	EC入	富岡 → 78
4.	1771-	51	E出	78 → 郡山	2.	1772-	52	EC入	郡山 → 78
5.	-				3.	-			
6.	-				4.	-			
7.	-				5.	-			
8.	-				6.	-			
その他 (分家=V, 引越=I, 行方不明=L, 追放=X, 理由不明=U, その他=O, 戻りは+C, 同伴は+W)									
1.	1747-	27	V	78 → 78A	2.	-			
2.	-				3.	-			
3.	-				4.	-			

(PR-1 71-10) PRNo 78-006

Source: Reitaku Archives, Population and Family History Project, Reitaku University

Note: ITS traces individual moves from the start and end of observation. The form includes N=name, B=birth, M=marriage, R=adoption, E=service, and Other (V=branching out; I=change of residence, etc.) with year, age, relationship to the household head, and place/households.

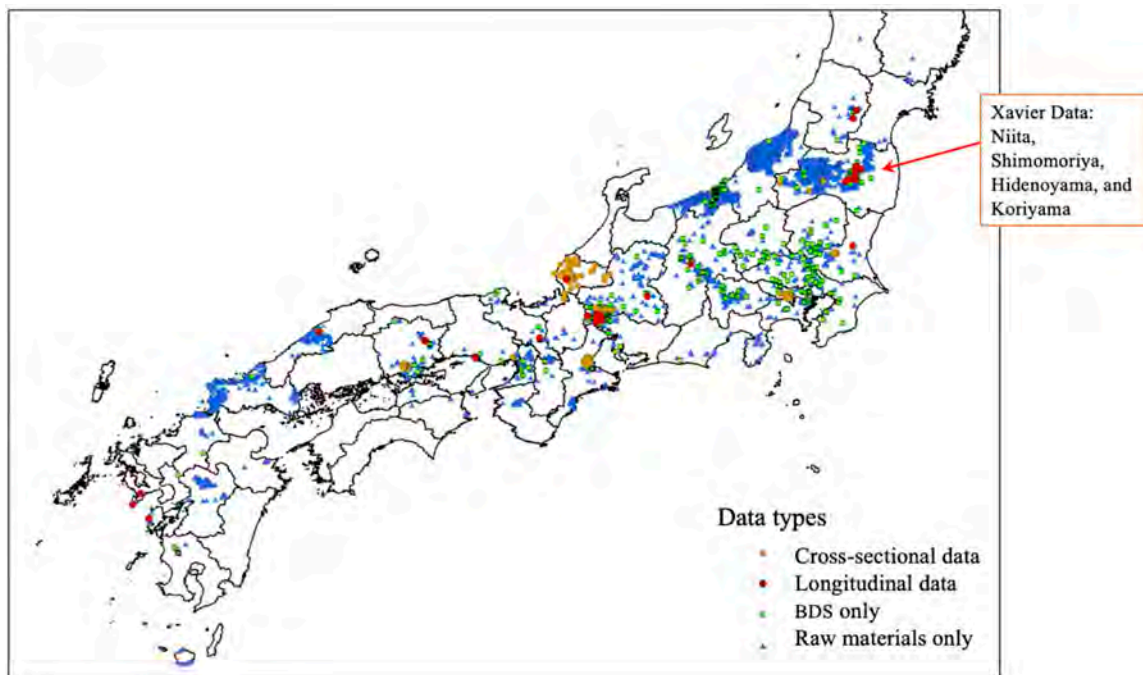
Figure 4 FRF (Family Reconstitution Form): A couple in household 112, Niita 1740–1781

分類	6	CF	00				6	姓名	仁井田	FRF No.	112-04						
夫名	仁井田	出生	1724-11	死亡	1781-11	結婚	1758-11	再婚	1758-11	宗派	112-04						
妻前名	下大和	結婚	1727-11	死亡	1710-12	結婚	1744-11	再婚	1758-11	宗派	112-04						
結婚	1740	終了	1781	終了	大和	出稼		入年		出先							
記号	年齢	継続	出生	順位	名前	性別	出生年月	母親	出生	死亡	死亡	結婚	結婚	結婚	宗派	備考	個人
	以下	1.5	0	1	仁井田	男	1742-10	16	2	1785-10	4						112-014
	16~20	5	1	2	伊平治	男	1747-6	2	3	-	-	1765	19	112-05		1788 1785	112-015
	21~25	5	1	3	伊平治	男	1752-7	2	5	-	-	1765	14	北本宮			112-016
	26~30	5	1	4	金作	男	1758-7	2	6	1741-8	24	1774	17	112-B-1			112-017
	31~35	5	1	5			-	-	-	-	-						
	36~40	5	0	6			-	-	-	-	-						
	41~45	5	0	7			-	-	-	-	-						
	46~50	5	0	8			-	-	-	-	-						
	51以上	4.5	0	9			-	-	-	-	-						
	合計	41.0	4	10			-	-	-	-	-						
		男	3	11			-	-	-	-	-						
		女	1	12			-	-	-	-	-						
	16~50	35.0	4	13			-	-	-	-	-						
		男	3	14			-	-	-	-	-						
		女	1	15			-	-	-	-	-						
身分		家格		職業			持高	1740	石	11.5	備考						

Source: Reitaku Archives, Population and Family History Project, Reitaku University.

Note: The form follows an internationally used format for family reconstitution. The top part of the panel contains information about husband and wife — name, place of birth, age at marriage, month/year of birth and death. No 1–4 in this example are their children with their name, sex, birth year/month, mother's age at birth, and their information on death and marriage. Information about births is summarized for every five years after marriage in the far left of the sheet (Hayami 2001, pp. 74–79).

Map 1 Xavier Data and Collections of Japanese Historical Demography Sources



Source: Reitaku Archives, Population and Family History Project, Reitaku University.

Note: See Table 1 for the explanation of the colored dots.

Table 1 *The different steps of database construction in relation to the map and figures*

Step	Work	Map/Figure
1	Making an inventory of all available sources	
2	Making a micro-film of the sources	Blue dots on Map 1
3	Making a Basic Data Sheet (BDS: systematic transcription of linked households for a maximum period of 25 years)	Green dots on Map 1, Figure 2
4	The identifiers to individuals are constructed	Figure 2 (Individual ID)
	Making an ITS for each individual	(Hayami's earlier trial) Figure 3
	Making a Family Reconstitution Form	(Hayami's earlier trial) Figure 4
5	Coding of BDS for Xavier data	Figure 5
6A	Computerizations of longitudinal data using BDS for villages with longitudinal records: Xavier data: based on coding sheets (Figure 5) Other data: based on BDS (Figure 2)	Red dots on Map 1
6B	Computerizations of cross-sectional data using BDS or other transcribed information sheets for villages with one-year records	Orange dots on Map 1

Notes: While steps 1–4 were applied to make forms from various sources that remain to digitize, steps 5 and 6A are specific to our construction of the Xavier Database, the subject of this article. Basic statistics of the Xavier Database are reported in Table 2.

As of now, Reitaku Archives include microfilmed or paper-copied documents of original SAC/NAC for about 1,870 villages/towns (Step 2). Among them, about 470 villages/towns (about 9960 village/town-years) transcribed in the format of BDS (Step 3). A meta-database of these materials has recently been made available online (limited use) to search the geographic location and type of historical records held at PFHP (Kurosu, 2020).

4 CONSTRUCTION OF THE XAVIER DATABASE

No official documentation exists as to what motivated Akira Hayami to initiate and how he pursued the construction of the Xavier Database. However, according to the detailed and yet complicated codebooks, as well as the memories of staff members who worked on the project over decades, the digitization was done with a series of trials and errors. What started in the 1980s remains an evolving process. In this section, we explain how the Xavier Database was transcribed and coded, turned into relational databases (DB2) for separate villages, and finally utilized for variable construction and analysis.

The Xavier data are based on household registers, both NAC and SAC in the northeast Japan. While this article mainly discusses four communities of Nihonmatsu domain (NACs), the Xavier Database also includes villages of the eastern-mountainous area of Fukushima⁵, Aizu, and the village Yanbe in the current Yamagata prefecture (SACs). Numerous studies used the BDS (step 4 in Table 1) of Aizu and Yanbe with some researchers' own data construction efforts (e.g., Hayami & Okada, 2005; Kinoshita, 2002; Okada, 2006). Also, since the NAC of Koriyama-shimo-machi, the northern town of Koriyama, has too many missing years for longitudinal research, we only include Koriyama-kami-machi, the southern town, in this article.

⁵ There is yet another village, Sasahara (NAC) in Nihonmatsu. As the size of the village was very small, it was conveniently used to test SQL runs but is not included in this article.

4.1 CODING AND DATA ENTRY

The first trial for the construction of a large longitudinal dataset started in the 1980s. Based on BDS, coders extracted information and organized them by handwriting into three paper forms, named A5, A6 and A7: household events and information (A5), individual age (A6), and individual events and life course (A7, see Figure 5). A5 had 10 household tables of codes and A7 had 17 individual tables of codes — including both time-constant and time-variant information as follows. For households, A5 codebook consists of the following tables: (A) entry and exit of households, (B) *de jure* or *de facto*, (C) head's information, (D) servants, (E) village official status, (F) size of household structure, (G) landholding, (H) livestock (horse, cattle), (I) ship ownership, (J) other. For individuals, A7 codebook consists of the following tables: (a) entry and exit of individuals, (b) birth, (c) death, (d) marriage (e) divorce, (f) adoption, (g) disowning (end of adoption), (h) change of residence, (i) household relationship, (j) change of name, (k) religion, (l) village official task, (m) migration, (n) other event with migration, (o) other event without migration, (p) other information (pregnant, illness, etc.), (q) other information that cannot be classified above.

Figure 5 Xavier data format A7 for the coding of individuals

A7 個人経歴シート

作成史料地名コード 0:5:10:8:10:9:4:10

史料初出の世帯コード	個人番号	出身	性別	対応番号	枚の内	枚目
0:0:0:3:10:0:0:0:0	0:1:4	1	1	: : : : :	/	/

父親コード	1 0:0:0:3:10:0:0:0:0	0:0:4	記入者: m. s.	INP	月日
母親コード	1 0:0:0:3:10:0:0:0:0	0:1:2	90年 月 日	照合	月日

年代	月	日	記載年	識別コード	記 入 一 点
9:9:9	9:9:9	9:9:9	7:5:2	410:110:3	1:1:9: : : : : : : : : : :
9:9:9	9:9:9	9:9:9	7:5:2	410:111:2	1:1: : : : : : : : : : :
9:9:9	9:9:9	9:9:9	8:4:0	410:110:4	1:2:8: : : : : : : : : :
9:9:9	9:9:9	9:9:9	8:4:0	410:111:7	0:9:9:0:0:1: : : : :
7:7:9	9:9:9	9:9:9	7:8:0	111:113:1	2:1:4:0:0:0:3:0:0:0:0:0:0:1:1: : : :
9:9:9	9:9:9	9:9:9	7:5:2	410:114:1	1:0:0:0:310:0:0:0:0:11:/:0:0:1: : : :
7:6:6	9:9:9	9:9:9	7:6:7	111:114:2	1:0:0:0:110:0:0:0:0:13:/:0:0:1: : : :
7:7:4	9:9:9	9:9:9	7:7:5	111:214:3	1:0:0:0:310:0:0:0:0:13:/:0:0:1: : : :
7:7:9	9:9:9	9:9:9	7:8:0	111:114:2	1:0:0:0:210:0:0:0:0:12:5:0:0:1: : : :
9:9:9	9:9:9	9:9:9	7:5:2	410:115:2	2:3:0:1: : : : : : : : : :
7:6:6	9:9:9	9:9:9	7:6:7	111:115:4	5:8:6:5:1: : : : : : : : :
7:7:4	9:9:9	9:9:9	7:7:5	111:215:4	5:3:0:1:1: : : : : : : :
7:7:9	9:9:9	9:9:9	7:8:0	111:115:4	5:3:2:1:1: : : : : : :
7:8:0	9:9:9	9:9:9	7:8:1	110:115:4	4:0:0:1:1: : : : : : :
8:3:1	9:9:9	9:9:9	8:3:2	110:115:4	4:1:2:3:1: : : : : : :
9:9:9	9:9:9	9:9:9	7:5:2	410:116:6	1:1:2: : : : : : : : :
8:1:2	9:9:9	9:9:9	8:1:3	110:115:4	4:1:0:3:1: : : : : :
9:9:7	9:9:9	9:9:9	7:5:2	410:116:1	1:8:2: : : : : : : :
7:8:2	9:9:9	9:9:9	7:8:3	510:216:1	2:1:1: : : : : : :
7:7:4	9:9:9	9:9:9	7:7:5	111:218:4	3:0:5:0:810:9:5:0:0:1:3:1:/10:3:0:0
7:7:5	9:9:9	9:9:9	7:7:6	510:219:1	6:5:0:5:018:1:0:7:010:2:0:1:1: : : :
7:2:6	9:9:9	9:9:9	7:7:7	510:219:1	6:5:0:5:018:1:1:1:010:2:0:2:1: : : :
7:7:8	9:9:9	9:9:9	7:7:7	112:118:3	4:0:5:0:811:1:1:0:10:2:4:1/10:3:0:0
7:8:1	9:9:9	9:9:9	7:8:2	111:218:4	3:0:5:0:810:9:1:0:010:3:3:1/10:3:0:0
7:8:4	9:9:9	9:9:9	7:8:5	510:219:1	6:5:0:5:018:0:9:7:010:2:0:3:1: : : :
7:8:6	9:9:9	9:9:9	7:8:7	510:219:1	6:5:0:5:018:0:9:1:010:2:0:4:1: : : :
7:8:8	9:9:9	9:9:9	7:8:9	510:219:5	6:5:4:0:111: : : : : :
7:9:2	9:9:9	9:9:9	7:9:3	510:219:1	6:5:0:5:018:1:1:1:010:2:0:5:1: : : :
7:9:3	9:9:9	9:9:9	7:9:4	510:219:1	6:3:0:5:018:0:6:2:010:2:0:6:1: : : :
9:9:4	9:9:9	9:9:9	7:9:5	112:118:3	4:0:5:0:810:6:2:0:0:4:4:1/10:2:0:0

Source: Reitaku Archives, Population and Family History Project, Reitaku University.
 Note: See text for details.

The codes in each table were elaborate and complicated. The origin of the Xavier data predated easy access to computers and even preceded the Japanese word processor for data entry (Ono, 1993). It was vital to convert the information written in BDS into numbers. Therefore, it required detailed codebooks. For example, there are more than 260 codes for the relationship to the household head A7(i) distinguishing kin relations, sex, blood or marriage/adoption relationship. Only trained coders could manually, using paper and pencil, convert detailed information of BDS into thousands of sheets that consist of sheer numbers. While the codes made data entry faster, the steps involved in coding cost enormous time and manpower, as well as an increasing number of errors as more steps are required (Ono, 1993).

The forms A5, A6 and A7 were then digitized manually using a special computer program. As of today, the coding of 7 villages and 2 towns has been done⁶ but the completion of the data entry was accomplished only for the four communities discussed in this article while the rest was, unfortunately, only partly completed.

4.2 THE XAVIER DATABASE AND THE EURASIA PROJECT

The database construction that started in Keio University in the early 1980s was carried over to Nichibunken (Kyoto), with Hayami's move in the late 1980s. It was at this time when the Xavier data were finally put to use. Yoshihiko Ono, associate professor at Nichibunken at the time, converted almost all the files into a relational database separating tables into time-constant and time-dependent tables using DB2 on an IBM platform. With the help of an international EAP member, George Alter, a group of EAP-J members⁷ headed by Yoshihiko Ono studied SQL and database management and started the construction of variables for international comparison. It was first tested on a small village, Sasahara, and then on the two villages of the Xavier Database (Shimomoriya and Niita) which became the basis for the Japanese contribution to the international collaboration of EAP (Tsuya & Kurosu, 2004, 2010c, 2014). Constructing variables for the EAP model required laborious processing. We used DB2 SQL to write long commands to construct each variable, export them to CSV files, and import and merge them in STATA for village-based analysis. This required advanced statistical and data management skills. It also required high levels of concentration and patience, as we had to go back-and-forth to the BDS and complicated codebooks whenever an inconsistency arose. The inconsistency could be due to the SQL program, data entry errors of coders, or data entry errors of BDS transcribers. It could even be due to the original documents — mistakes of village officials in Tokugawa period. With all these laborious and time-consuming processes, we were not able to add more than two villages during the entire EAP international collaboration (Kurosu, 2016). In the late 2000s, however, we started to add another village and town from the Xavier Database to our analysis: Hidenoyama and Koriyama.

4.3 THE XAVIER DATABASE: POPULATION AND SETTINGS

Table 2 shows a summary of the database for the four communities as well as the information available in those communities. Altogether they cover the span of 162 years with a relatively small number of missing years. In addition to its long coverage, the Xavier Database provides detailed information about demographic characteristics and reason for individual "entrance" under observation (due to birth or immigration) and the month/year of death and other "exits". This is the case even for Koriyama despite the heavy in- and out-migration of the town. Socioeconomic indicators, which are not often recorded in SAC in other places, are also abundant. The two Xavier villages (Niita and Shimomoriya), in particular, are among the best quality historical population panel data available in East Asia (Dong, Campbell, Kurosu, Yang, & Lee, 2015a).

That said, we need to be careful about one problem inherent in "annual" household registers: the omission of events that happened between registers. NACs of these communities were enumerated annually at the beginning of the third lunar month⁸, an important point to bear in mind for demographic calculations. The timing of marriage, for example, cannot be determined as clearly as many European

6 This includes Shimomoriya, Niita, Hidenoyama, Sasahara, Koriyama-kami, Koriyama-shimo in Nihonmatsu which are complete, and Kuwahara, Kanaizawa, Ishibuse in Aizu, and Yanbe in Murayama. The continuation of data entry unfortunately was disrupted due to Hayami's move from Keio University to the International Research Center for Japanese Studies (Nichibunken).

7 Satomi Kurosu and Hideki Nakazato worked intensively with Ono. Noriko O. Tsuya was instrumental in defining demographic variables proposed in the EAP model for the context of Japan.

8 The timing of NAC/SAC varied by region.

parish registers. If one comes into a household with an annotation of *enduke* (marriage), the person and his/her partner married sometime between the two register years. An even more serious problem for certain demographic analysis is the omission due to infant death. Those who were born and died between the two registers may not be recorded.

Map 2 shows the location of the four communities. Shimomoriya and Niita were almost totally agricultural with an annual average population size of about 300 and 450, respectively. Hidenoyama, located about 3 km from the center of the growing market town of Koriyama, became a suburb to this town with an annual average population of about 280. Within the 160 years of observation, however, the area has gone through major famines in the 1780s and 1830s, as well as various epidemics and local disasters. The population trend in Figure 6 reflects the hardship the villages went through. In particular, the two agricultural villages, Shimomoriya and Niita, started to lose their population before the Tenmei famine in the 1780s and did not recover until the 19th century. Such constant decline is not observed in Hidenoyama. As for Koriyama, being a post town⁹ with diverse economic activities, the local population increased over the years from 800 to about 2,600 inhabitants while commercial sectors developed. Koriyama was formally designated as a town (*machi*) in 1824 and served as the economic as well as the political center of Asaka County. While populations of neighboring villages stagnated, Koriyama experienced a stable population growth. Although it was affected by famines, the population of Koriyama soon recovered because of both in-migration and natural increase. Thus, the Xavier Database shows different population trends between Koriyama and neighboring villages, suggesting an interesting contrast of demographic dynamics between the rural and urban communities.

Table 2 Summary and information of the Xavier Database

Village/Town	Observation period	Years of missing registers	Person - years	Unique individuals
Niita	1720–1870	4	74,099	4,076
Shimomoriya	1716–1869	9	53,628	2,468
Hidenoyama	1708–1870	35	40,036	3,046
Koriyama	1729–1870	18	219,503	18,515
Total			346,700	28,105

Information available in four communities			
Demographic characteristics and events	Inter-personal relations	Socioeconomic indicators	Other information
Age, sex*, birth, death, marriage, divorce, adoption, service, labor-migration	Relation to the household head, conjugal, sibling, parent-child, multi-generational	Land-owner or landless peasants, village officials, household landholding, number of horses	Land rent/lease (Niita and Shimomoriya only)

Note: *Sex is inferred based on relationship to the household head and/or name.

9 Post towns, *Shukuba-machi* in Japanese, were constructed along the major routes/streets in the Tokugawa period. They provided lodgings for public officials, who were forced to periodically travel between their domain and Edo with their vassals; as well as rest for travelers, who were observed more frequently as traveling became more popular throughout the country during this period. Commercial sectors also developed in these towns and catered to the needs of commoners and neighboring villagers (Kurosu, Takahashi, & Dong, 2017).

Map 2 Present-day Fukushima prefecture — location of Niita, Shimomoriya, Hidenoyama and Koriyama

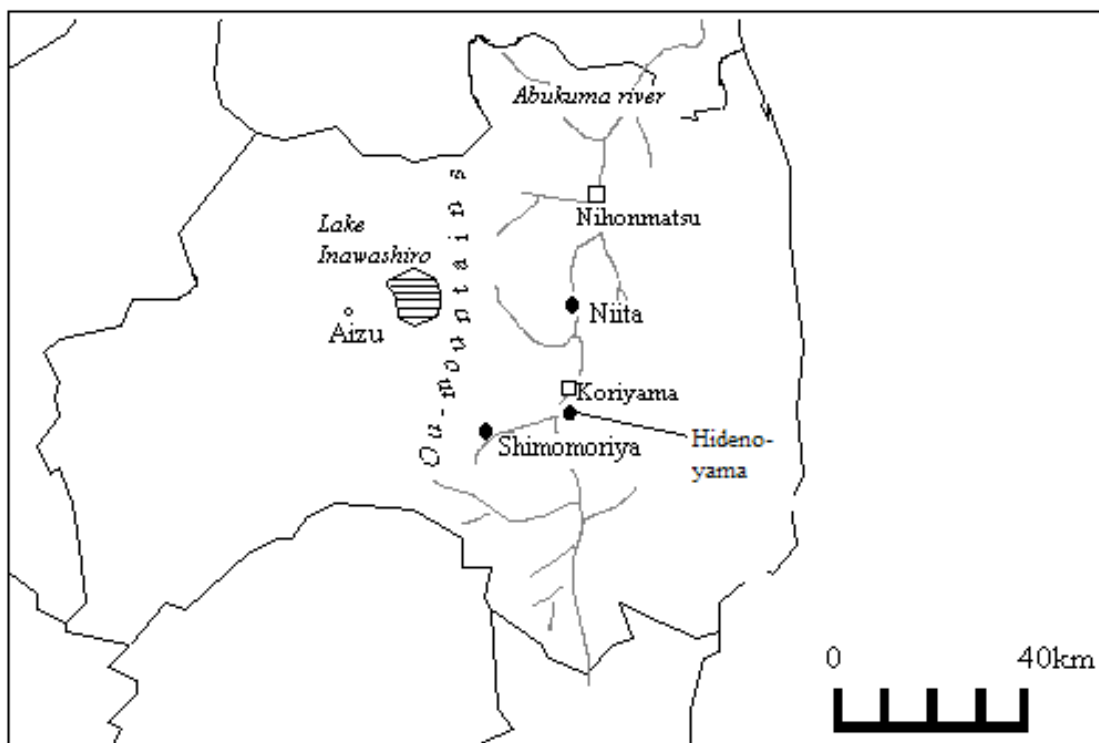
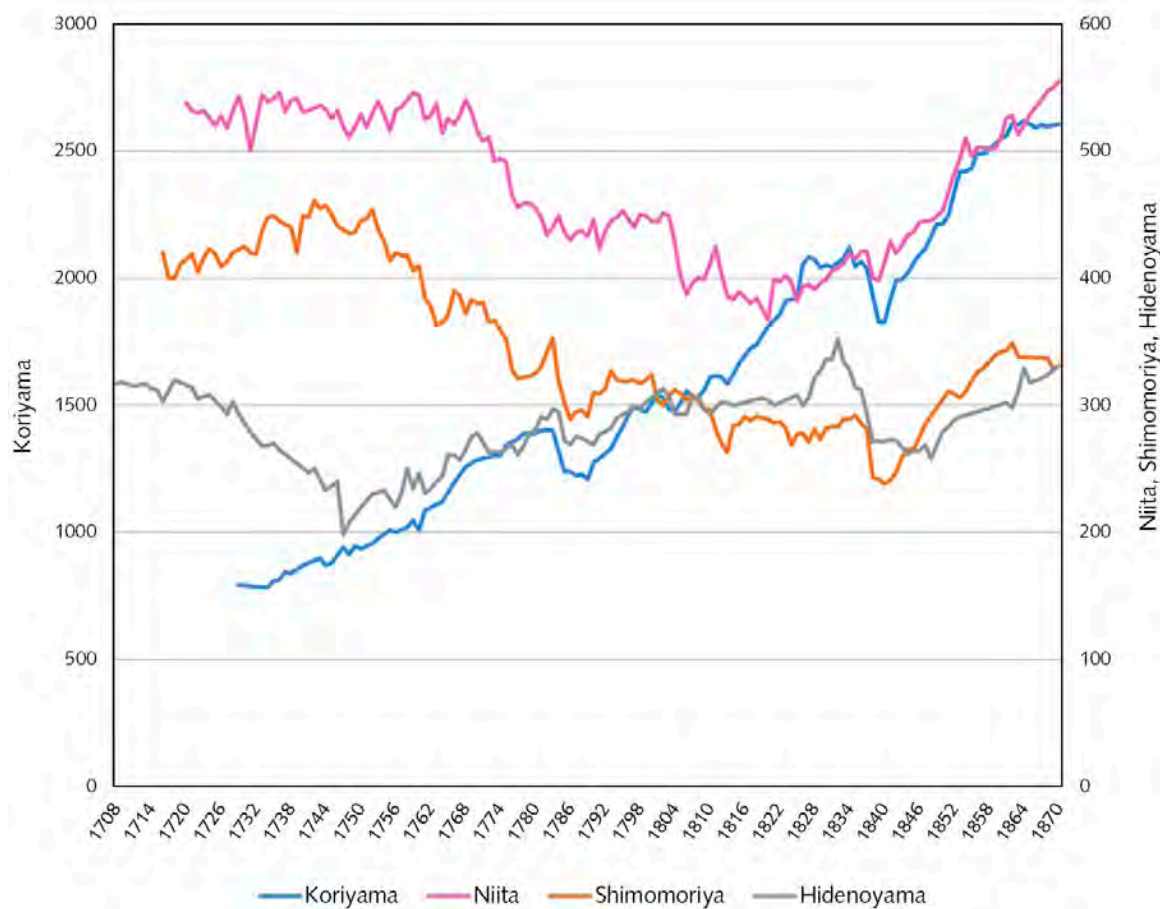


Figure 6 Population trends of Niita, Shimomoriya, Hidenoyama, and Koriyama



5 RESEARCH WITH THE XAVIER DATABASE

Many advances in the use of the data occurred during the Japanese Eurasia Project. Flat files for analysis based on the two villages, Niita and Shimomoriya, were first made specifically for the international comparison of mortality in EAP (Bengtsson, Campbell, Lee, et al., 2004), and successively for marriage and reproduction (Lundh, Kurosu, et al., 2014; Tsuya & Kurosu, 2010c). Japanese contributions to these volumes were based on Niita and Shimomoriya (Tsuya & Kurosu, 2004, 2010c, 2014). Modifying the models used in the international collaboration of EAP, Tsuya and Kurosu produced more detailed studies of mortality and reproduction (Tsuya & Kurosu, 2005, 2010b). These studies brought new findings about how individual demographic behaviors were influenced by socioeconomic status of households (indicated by landholdings), household context (measured by the presence of co-residing kin), and status in the household (head or closeness to household-head). Further, applying the EAP models but going beyond the topics covered by the EAP international comparison, specific topics for Japan were investigated utilizing the information available in the Xavier Database. These studies emphasized the hierarchy of gender and closeness to household-head in the stem family system; for example, the differential stress caused by the deaths of heads vs. fathers on women and children (Tsuya & Kurosu, 2002), household context and migration and survivorship (Tsuya & Kurosu, 2010a, 2013), leaving home, remarriage, divorce, and adoption (Kurosu, 2004, 2007, 2011, 2013). These studies used event history analysis and revealed how the principle of the stem family mediated individual as well as household choices.

Later on, similar models for mortality and marriage were tested with a larger sample by including the village of Hidenoyama (Tsuya & Kurosu, 2017, 2020). Further, including the town of Koriyama, we started to systematically compare patterns of marriage between village and town populations (Kurosu, Takahashi, & Dong, 2017). The EAP model was also tested with more rigorous statistical analysis to investigate the association between types of post-marital residence and reproduction (Dong & Kurosu, 2017). These studies demonstrated "northeastern characteristics" of marriage and reproduction influenced by the norm of "primogenitor" (succession by first child regardless of sex) that distinguished the area, going even beyond rural and urban differentials.

Researchers who shared knowledge and methods for working with the Xavier Database flourished in the Japanese Eurasia Project. They used various methods (both descriptive and statistical) to empirically tackle issues of family history which had been confined to institutional or household-based approaches. Comprehensive studies of Koriyama and Niita appeared in two Ph.D. dissertations using the Xavier Database and complementary local historical records (Hirai, 2008; Takahashi, 2005)¹⁰. Takahashi (2005) was one of the first who systematically studied a booming town, Koriyama, and indicated the importance of understanding migration as well as domanical population policy in the demographic model. The urban graveyard theory was examined but found not to be applicable for this town. Hirai (2008) studied the dataset of Niita from a comprehensive approach to individual life courses and households and found that households became more resilient and stable towards the 19th century.

Xavier data allowed a dynamic approach from individual life course perspectives providing new findings and detailed accounts for marriage, adoption, and their outcomes (Hirai, 2006a, 2015; Kurosu, 1997, 1998, 2007, 2009, 2011, 2012b; Takahashi, 2012), living arrangements (Hirai, 1998; Nakazato, 1999, 2009; Ochiai, 2001, 2006c), leaving home (Kurosu, 2004), service and labor migration (Nagata, 2001, 2004; Ochiai, 2002; Takahashi, 2019), and name changing (Nagata, 1999, 2009). These studies found how individual life courses were stratified by the rule of stem family organization, i.e., the overall aim of the continuity of households that shaped individual demographic events in order to overcome various demographic and economic constraints. The Japanese historical family was further examined in an international comparison aimed at comparing stem family societies in Europe and Japan (Fauve-Chamoux & Ochiai (Eds.), 2009). Individual-level studies of marriage, name-changing and living-arrangement based on the Xavier Database (Kurosu, 2009; Nagata, 2009; Nakazato, 2009) reveal the dynamics of individual life courses within the stem family system.

The longitudinal data of the Xavier Database also offered dynamic approaches to "old" questions about household structure, continuity, and succession. Okada applied the studies of household cycles to Nihonmatsu villages and asserted a modified version of Hammel-Laslett in order to show the clear

¹⁰ Yet another dissertation by Okada (2006) demonstrated the developmental cycle of households in Aizu domain proposing a modified Hammel-Laslett model. She did not directly use the Xavier Database but BDS of Aizu villages from which the Xavier Database was constructed.

cyclical change of Japanese households as well as succession strategies (1998, 2000, 2002). The stem family orientation of the peasant family became clear, particularly among land-owning peasants. Hirai identified characteristics of the continuity of households spanning over a century (e.g., 2003, 2006b).

This series of studies on individual life courses and household cycles written in both English and Japanese became the core achievements of the Japanese Eurasia Project (Ochiai, 2006b) and affected the fields of historical demography and family history and sociology in various ways. First, the novelty of findings sustained by detailed and statistically sound analysis painted clear and fresh pictures of commoners' life courses and families in the 18–19th centuries. For example, universal marriage was followed by flexible divorce and remarriage. Early marriage did not necessarily bring an early start of childbearing. Fertility was extremely low because of the practice of sex-selective infanticide, and adoption compensated for demographic constraints when sons or daughters were lacking. These demographic behaviors were in accordance with peasants' strategy of keeping the stem family intact. These new findings were surprises to family sociologists influenced by the western tradition of family modernization, which emphasizes monolithic and developmental changes. They also made demographers and historians reconsider the importance of households and the stem family system in understanding the contrasting population dynamics in Japan compared with other countries (e.g., Ochiai, 2009; Saito, 1998, 2000).

Studies based on Xavier data opened possibilities for broadening research into two directions: longitudinal and regional approaches. First, within about 160 years of observation, a general improvement of climate and development of cash crops as well as sericultural industries in the region boosted development of the regional economy. Various improvements in demographic behaviors are observed, particularly related to females: increase in the female age of marriage (Kurosu, 2012a; Kurosu, Takahashi, & Dong, 2017; Tsuya & Kurosu, 2014), decline of female infant mortality (Tsuya & Kurosu, 2004), and increase of reproduction (Tsuya & Kurosu, 2010b, 2010c). Also, homogenization of age at marriage and shortening of length between marriage and divorce were associated with increasing household size and complexity making households more stable towards the end of Tokugawa period. Hirai calls the process "emergence" of Japanese traditional family that emphasizes its continuity and the stem family household (Hirai, 2003, 2006a, 2006b, 2008).

Second, the characteristics of northeastern population and family were clarified empirically in the comparative framework of regional variation in Japan (e.g., Kurosu, Tsuya, & Hamano, 1999; Ochiai (Ed.), 2006a; Okada & Kurosu, 1998; Takahashi & Kurosu, 2020). While fertility was generally low in early modern Japan, it was very low in the northeast (Takahashi, 2005; Tsuya & Kurosu, 2010b, 2010c). The survival rate tends to be lower than other studied areas of Japan (Tsuya & Kurosu, 2004). Marriage was early, universal, and more flexible than in the rest of Japan (Kurosu, Tsuya, & Hamano, 1999). Labor migration before marriage delayed marriage in central Japan, but in the northeast migration took place after marriage, making marriage a safety net for households to have them return for sure after service (Nagata, 2001, 2004). Researchers argue that the strategy was intended to increase the working-age population in households to overcome environmental hardships in the northeastern region (Hayami, 2009; Hayami & Kurosu, 2001). These arguments, however, are only based on limited village studies of the northeast, Nishijo in central, and Nomo in the southwestern Japan¹¹, and await further investigation.

6 CONCLUSION

This article introduced one of the major databases of Japanese historical demography, called the Xavier Database, which includes individual-level longitudinal data of populations in three villages and one town of the current Fukushima prefecture between 1708 and 1870. We discussed the long and complicated process used by the founder of Japanese historical demography, Akira Hayami, and his colleagues to collect, transcribe, code, and finally put the content of the historical population registers into a database. We also reviewed the content and studies that flourished domestically and internationally using these data in the last two decades.

11 Databases for Nishijo (NOBI) and Nomo (NOMO) were constructed during the Japanese Eurasia Project, and also gave opportunities for developing sociological and/or statistical investigations using historical records (e.g., Nakajima, 2016; Nakazato, 2004).

During the last decade, new attempts are being made into at least four directions. First, the two villages of the Xavier Database, Niita and Shimomoriya, are harmonized and used for several comparative studies with other East Asian historical populations from northeastern China, Taiwan, and Korea — a collaborative effort sometimes referred to as the East Asia Project or EAP II (Dong et al., 2015a). The standardization and harmonization of different East Asian population datasets, an effort led by Hao Dong in collaboration with each relevant research team, have enabled detailed, systematic comparative studies about similarities and differences in population dynamics and family histories within East Asia (e.g., Dong, 2016; Dong, Campbell, Kurosu, & Lee, 2015b; Dong, Kurosu, & Lee, 2019; Dong, Manfredini, Kurosu, Yang, & Lee, 2017). Second, although still being at the experimental stage, comparative studies of early modern and post-modern Japanese families reveal some interesting resiliency or continuity of Japanese family. These studies bridge Tokugawa and contemporary Japan applying the same model for divorce and marriage to Xavier data and contemporary survey data (Kurosu & Kato, 2018; Tsuya & Kurosu, 2019). Third, there is a new effort coding and adding more information to the original Xavier Database from records of migration and land-lease. In the last few years, we identified geographic locations of 5,000 migration records from the Xavier Database. We can track where migrants came from and went to and the reasons for migration such as service, marriage, or adoption. This will add a spatial dimension to the longitudinal analysis (Kurosu, 2020; Kurosu, Takahashi, & Nagaoka, 2017; Nagaoka, Kurosu, & Takahashi, 2020). Other unique information of the Xavier Database not used until now is land lease/rent and livestock. We added more details of the land transactions (from whom, to whom) for the village of Niita. This has started to show us how land transactions were associated with social mobility and equality as well as demographic patterns of the villages (Arimoto & Kurosu, 2020). We are also constructing a database on livestock based on the records of horses and cattle in BDS. This will make additional useful information about agricultural conditions and the peasant economy related to population dynamics (Takahashi, 2018). Finally, efforts led by Satomi Kurosu with the cooperation of Hao Dong have been undertaken at PFHP to integrate various data collections and sources to advance larger-scale, longitudinal research of family behavior and population history in early modern Japan¹². The comparative and interdisciplinary approach applied to records of thousands of lives promises new understanding of our history and the resilience of people to socioeconomic and environmental changes.

As the foundation for all, the Xavier Database still inspires us with interesting research topics and possibilities going beyond the field of historical demography and the family. We are extremely grateful for the inexhaustible research opportunities provided by the database.

Thank you, Francisco Xavier! Thank you, Akira Hayami!!

ACKNOWLEDGEMENTS

We are grateful to Akira Hayami (1928–2019) for leaving us amazing sources and datasets, and for his leadership, insights, and guidance in the studies of historical demography. Part of this article is based on the paper we co-authored with Hayami and presented at the 2019 Annual Meeting of the Social Science History Association in Chicago, IL, "Constructing Individual-Level Longitudinal Data for Japanese Historical Population: Challenges and Opportunities". We would like to thank Saeko Narimatsu and all the current and former project members who contributed to collecting, transcribing, and organizing old documents as well as entering and cleaning the data sources. We would like to thank Shoko Hirai, Hideki Nakazato, Aoi Okada, and Atsushi Nagaoka, researchers who worked with Xavier data; Toshiko Mochida, Hirofumi Ohnuma, and Yuriko Kikuchi, research assistants at PFHP for preparing this draft.

12 This project is supported by "Social and geographic mobility and life course: Studies using multi-generational panel data" Grants in Aid for Scientific Research (B), Japan Society for the Promotion of Science (KAKEN 19H01569).

REFERENCES

- Arimoto, Y., & Kurosu, S. (2020). Tokugawa Nihon noson no shisan-bunpai: Nihonmatu-han Niitamura (1720–1870) wo jireini [Asset allocation in rural Tokugawa Japan: The case of Niita village of Nihonmatsu domain, 1720–1870]. *Economic Review*, 71(3), 237–258.
- Bengtsson, T., Campbell, C., Lee, J. Z., et al. (2004). *Life under pressure: Mortality and living standards in Europe and Asia, 1700–1900*. Cambridge, MA: MIT Press.
- Cornell, L. L. (1981). *Peasant family and inheritance in a Japanese community: 1671–1980: An anthropological analysis of local population registers* (Doctoral dissertation). Johns Hopkins University, Baltimore, MD.
- Cornell, L., & Hayami, A. (1986). The Shumon Aratame Cho: Japan's population registers. *Journal of Family History*, 11(4), 311–328. doi: [10.1177/036319908601100401](https://doi.org/10.1177/036319908601100401)
- Dong, H. (2016). *Patriarchy, family system and kin effects on individual demographic behavior throughout the life course: East Asia, 1678–1945* (Doctoral dissertation). The Hong Kong University of Science and Technology, China.
- Dong, H., Campbell, C., Kurosu, S., Yang, W., & Lee, J. Z. (2015a). New sources for comparative social science: Historical population panel data from East Asia. *Demography*, 52(3), 1061–1088. doi: [10.1007/s13524-015-0397-y](https://doi.org/10.1007/s13524-015-0397-y)
- Dong, H., Campbell, C., Kurosu, S., & Lee, J. Z. (2015b). Household context and individual departure: The case of escape in three "unfree" East Asian populations, 1700–1900. *Chinese Journal of Sociology*, 1(4), 515–539. doi: [10.1177/2057150X15614547](https://doi.org/10.1177/2057150X15614547)
- Dong, H., & Kurosu, S. (2017). Postmarital residence and child sex selection: Evidence from northeastern Japan, 1716–1870. *Demographic Research*, 37, 1383–1412. doi: [10.4054/DemRes.2017.37.43](https://doi.org/10.4054/DemRes.2017.37.43)
- Dong, H., Kurosu, S., & Lee, J. Z. (2019, November). *The making of missing girls: Evidence from northeast China and Japan, 1708–1909*. Presentation at the annual meeting of Social Science History Association, Chicago, IL.
- Dong, H., Manfredini, M., Kurosu, S., Yang, W., & Lee, J. Z. (2017). Kin and birth order effects on male child mortality: Three East Asian populations, 1716–1945. *Evolution of Human Behavior*, 38(2), 208–216. doi: [10.1016/j.evolhumbehav.2016.10.001](https://doi.org/10.1016/j.evolhumbehav.2016.10.001)
- Drixler, F. (2013). *Mabiki: Infanticide and population growth in eastern Japan, 1660–1950*. Oakland, CA: University of California Press.
- Fauve-Chamoux, A., & Ochiai, E. (Eds.). (2009). *The stem family in Eurasian perspective: Revisiting household societies, 17th–20th centuries*. Bern: Peter Lang.
- Hayami, A. (1973). *Kinsei noson no rekishi-jinko-gaku-teki kenkyu shinshu Suwa-chiho no Shumon Aratame-cho bunseki* [Historical demographic study of early modern agricultural villages: Analysis of Shumon-Aratame-cho in Suwa region, Shinshu]. Tokyo: Toyo-Keizai-Shimpo-sha.
- Hayami, A. (1979). Thankyou, Francisco Xavier: An essay in the use of micro-data for historical demography of Tokugawa Japan. *Keio Economic Studies*, XVI (1–2), 65–81. Retrieved from https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AA00260492-19790001-0065
- Hayami, A. (1992). *Kinsei Nobi-chiho no jinko* [Population of early modern Nobi area]. Tokyo: Sobunsha.
- Hayami, A. (2009). *Population, family and society in pre-modern Japan (Collected papers of twentieth-century Japanese writers on Japan)*. Kent: Global Oriental.
- Hayami, A. (2001). *The historical demography of pre-modern Japan*. Tokyo: University of Tokyo Press. (Originally published in Japanese by Iwanami Shoten Publishers, 1997)
- Hayami, A., & Kurosu, S. (2001). Regional diversity in demographic and family patterns in preindustrial Japan. *Journal of Japanese Studies*, 27(2), 295–321. doi: [10.2307/3591968](https://doi.org/10.2307/3591968)
- Hayami, A., & Okada, A. (2005). Population and household dynamics: A mountainous district in northern Japan in the Shûmon Aratame Chô of Aizu, 1750–1850. *The History of the Family*, 10(3), 195–229. doi: [10.1016/j.hisfam.2005.03.007](https://doi.org/10.1016/j.hisfam.2005.03.007)
- Hirai, S. (1998). Kinsei niokeru yome-shutome no kyoju-keitai: Nihonmatu-han Niita-mura no jirei yori [Living arrangements between wives and mothers-in-law in rural northeastern Japan, 1720–1870. *Kazoku-Kenkyu-Sosho* [Nara Journal of Family Studies], 4, 3–20.
- Hirai, S. (2003). Kinsei tohoku-noson niokeru "ie" [Emergence of the "ie" or the Japanese "traditional" family in early modern Japan: A historical-demographic analysis]. *Soshioroji* [Sociology], 47(3), 3–18.
- Hirai, S. (2006a). Kekkon no kinshitsu-ka to "ie" no kakuritsu: Tohoku-noson no baai [Homogenization of marriage and the establishment of the "ie": A case of northeastern farming village]. In E. Ochiai (Ed.), *Tokugawa Nihon no raifu kosu* [Life courses of Tokugawa Japan]. Kyoto: Minerva-shobo.

- Hirai, S. (2006b). *Ie no kakuritsu to kasan no keisho: Mutsu-no-kuni Adachi-gun Niita-mura no jirei* [The establishment of the family and the succession of family property: A case study of Niita, Adachi County, Mutsu province]. In E. Ochiai (Ed.), *Tokugawa no kazoku to chiiki-sei* [Family and regionality of Tokugawa Japan]. Kyoto: Minerva-shobo.
- Hirai, S. (2008). *Nihon no kazoku to raifu kosu "ie" seisei no Rekishi-shakai-gaku* [Family and life course in Japan: Historical sociology of the emergence of the "ie"] (Doctoral dissertation). Kyoto: Minerva-shobo.
- Hirai, S. (2011). Tohoku-nihon niokeru ie no rekishi-jinko-gaku-teki bunseki [A historical demographic analysis of households in northeastern Japan: Focus on population trends in the eighteenth and nineteenth centuries]. In K. Kasaya (Ed.), *18-seiki Nihon no bunka-jokyo to kokusaikankyo* [Japan in the eighteenth century: Cultural conditions and international environment] (pp. 215–231). Kyoto: Shibunkaku Shuppan.
- Hirai, S. (2013). Kinsei-sonraku niokeru ie no henyō [Changes in the ie system in early modern Japanese villages]. *Shakaigaku Zasshi* [Sociological Review of Kobe University], 30, 78–90.
- Hirai, S. (2015). Tohoku-noson niokeru kekkon pataan no henyō: 18–19 seiki no rekishi-jinko-gaku-teki bunseki [Transformation of marriage patterns in rural northeastern villages: A historical demographic analysis of the 18th and 19th centuries]. In K. Kasaya (Ed.), *Tokugawa-shakai to Nihon no kindai-ka* [Tokugawa society and the modernization of Japan] (pp. 407–423). Kyoto: Shibunkaku Shuppan.
- Kinoshita, F. (2002). *Kindaika-izen no Nihon no jinko to kazoku: Ushinawareta sekai karano tegami* [Population and family before Japanese modernization: Letter from the lost world]. Kyoto: Minerva-shobo.
- Kurosu, S. (1997). Adoption as an heirship strategy? A case from a northeastern village in pre-industrial Japan. *Japan Review*, 9, 171–189. Available from <http://www.jstor.org/stable/25791007>
- Kurosu, S. (1998). Long way to headship, short way to retirement: Adopted sons in a northeastern village in pre-industrial Japan. *The History of the Family*, 3(4), 393–410. doi: [10.1016/S1081-602X\(99\)80254-1](https://doi.org/10.1016/S1081-602X(99)80254-1)
- Kurosu, S. (2004). Who leaves home and why: Daughters and sons in two northeastern villages, 1716–1870. In F. van Poppel, M. Oris, & J. Z. Lee (Eds.), *The Road to independence: Leaving home in western and eastern societies, 16th–20th centuries* (pp. 243–271). Bern: Peter Lang.
- Kurosu, S. (2007). Remarriage in a stem family system in early modern Japan. *Continuity and Change*, 22(3), 429–458. doi: [10.1017/S026841600700642X](https://doi.org/10.1017/S026841600700642X)
- Kurosu, S. (2008). Filling gaps in Japanese historical demography: Marriage, fertility, and households in nineteenth-century rural Japan. *Sungkyun Journal of East Asian Studies*, 18(1), 43–70. Retrieved from <https://sjeas.skku.edu/#/search/detail/4163?offset=1>
- Kurosu, S. (2009). Marriage, divorce and remarriage in a stem family system: Women in two northeastern Japanese villages, 1716–1870. In A. Fauve-Chamoux & E. Ochiai (Eds.), *The stem family in Eurasian perspective: Revisiting household societies, 17th–20th* (pp. 327–344). Bern: Peter Lang.
- Kurosu, S. (2011). Divorce in early modern rural Japan: Household and individual life course in northeastern villages, 1716–1870. *Journal of Family History*, 36(2), 118–141. doi: [10.1177/0363199011398428](https://doi.org/10.1177/0363199011398428)
- Kurosu, S. (2012a). Mukotori-kon to yomeiri-kon: Tohoku-noson niokeru joshi no kekkon to raifu kosu [Uxorilocal vs. virilocal marriage: Marriage and life-course among women in northeastern rural villages]. In S. Kurosu (Ed.), *Rekishi-jinko-gaku karamita kekkon, rikon, saikon* [Marriage, divorce and remarriage from the perspective of historical demography] (pp. 57–79). Kashiwa: Reitaku University Press.
- Kurosu, S. (Ed.) (2012b). *Rekishi-jinko-gaku karamita kekkon, rikon, saikon* [Marriage, divorce and remarriage from the perspective of historical demography]. Kashiwa: Reitaku University Press.
- Kurosu, S. (2013). Adoption and family reproduction in early modern Japan. *The Economic Review*, 64(1), 1–12. doi: [10.15057/25877](https://doi.org/10.15057/25877)
- Kurosu, S. (2016). Historical demography going 'glocal': Eurasia project and Japan. In K. Matthijs, S. Hin, J. Kok, & H. Matsuo (Eds.), *The future of historical demography: Upside down and inside out* (pp. 60–62). Leuven/Den Haag: Acco. Retrieved from <https://soc.kuleuven.be/ceso/fapos/publications/the-future-of-historical-demography-upside-down-and-inside-out>
- Kurosu, S. (2020). Reitaku archives no kinsei jinko-keizai shiryō-Hayami Akira-shi kizo shiryō no meta-database kochiku [Reitaku archives and early modern demographic and economic data: Construction of meta-database for Hayami collection]. *Gengo to Bunmei* [Language and Civilization], 18(2), 27–38.

- Kurosu, S., & Kato, A. (2018, April). *Socioeconomic factors of divorce: A comparative analysis of early modern and contemporary Japan*. Presentation at the annual meeting of Population Association of America, Denver, CO.
- Kurosu, S., Takahashi, M., & Dong, H. (2017). Marriage, household context and socioeconomic differentials: Evidence from a northeastern town in Japan, 1716–1870. *Essays in Economic and Business History*, 35(1), 239–263. Retrieved from <https://www.ebhsoc.org/journal/index.php/ebhs/article/view/49>
- Kurosu, S., Takahashi, M., & Nagaoka, A. (2017). Xavier data kara hukugen suru ido hisutori: Kinsei shomin no jinko ido shiryō [Migration history reconstructed based on Xavier data]. *Gengo to Bunmei* [Language and Civilization], 15, 139–150. Retrieved from https://reitaku.repo.nii.ac.jp/?action=pages_view_main&active_action=repository_view_main_item_detail&item_id=992&item_no=1&page_id=13&block_id=29
- Kurosu, S., Tsuya, N. O., & Hamano, K. (1999). Regional differentials in the patterns of first marriage in the latter half of Tokugawa Japan. *Keio Economic Studies*, 36(1), 13–38. Retrieved from https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AA00260492-19990001-0013
- Lundh, C., Kurosu, S., et al. (2014). *Similarity in difference: Marriage in Europe and Asia, 1700–1900*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262027946.001.0001
- Nagaoka, A., Kurosu, S., & Takahashi, M. (2020). Kinsei tohoku niokeru Mutsu-no-kuni Nihonmatsuhan choson no jinko ido no kukanteki hirogari [Migration and geographical spread in early-modern Nihonmatsu in the northeastern region]. *Gengo to Bunmei* [Language and Civilization], 18(2), 17–6. Retrieved from https://reitaku.repo.nii.ac.jp/?action=pages_view_main&active_action=repository_view_main_item_detail&item_id=1263&item_no=1&page_id=13&block_id=29
- Nagata, M. L. (1999). Balancing family strategies with individual choice: Name changing in early modern Japan. *Japan Review*, 11, 145–166. Available from <http://www.jstor.org/stable/25791040>
- Nagata, M. L. (2001). Labor migration, family and community in early modern Japan. In P. Sharpe (Ed.), *Women, gender and labour migration: Historical and cultural perspectives* (pp. 60–84). London: Routledge.
- Nagata, M. L. (2004). Leaving the village for labor migration in early modern Japan. In F. van Poppel, M. Oris, & J. Z. Lee (Eds.), *The Road to independence: Leaving home in western and eastern societies, 16th–20th centuries*. Bern: Peter Lang.
- Nagata, M. L. (2009). Name changing patterns and the stem family in early modern Japan: Shimomoriya. In A. Fauve-Chamoux & E. Ochiai (Eds.), *The stem family in Eurasian perspective: Revisiting household societies, 17th–20th* (pp. 361–378). Bern: Peter Lang.
- Nakajima, M. (2016). *Kinsei seinan-kaison no kazoku to chiikisei: Rekishi-jinko-gaku kara kindai no hajimari wo tou* [Family and community in the early-modern southwestern fishing village: Questioning the beginning of modernization from historical demography]. Kyoto: Minerva-shobo.
- Nakazato, H. (1999). Kinsei tohoku noson niokeru koreisha no kyojyukeitai: Oyako ichiji dokyo no pataan [Living arrangements of the elderly in early modern rural Japan: Transitions between presence and absence of coresiding children]. *Kazoku-Kenkyu-Sosho* [Nara Journal of Family Studies], 5, 3–20.
- Nakazato, H. (2004). Rireishonaru database niyoru teikei deta no sakusei: Shumon-Aratamae-Cho no tokei bunseki no tameni [Creating flat-files for analysis using relational database: For the statistical analysis of Shumon-Aratame-Cho]. *Riron to Hoho* [Theory and method], 19(2), 197–212. doi: 10.11218/ojjams.19.197
- Nakazato, H. (2009). Transitions in living arrangements over the life course: Aging in a rural village in Japan, 1716–1869. In A. Fauve-Chamoux & E. Ochiai (Eds.), *The stem family in Eurasian perspective: Revisiting household societies, 17th–20th* (pp. 345–360). Bern: Peter Lang.
- Narimatsu, S. (1985). *Kinsei tohoku noson no hitobito: Oshu Asaka-gun Shimomoriya-mura* [People in a northeastern agricultural village in early modern Japan: The village of Shimomoriya, Asaka county, Ou region]. Kyoto: Minerva-shobo.
- Narimatsu, S. (1992). *Edo-jidai no Tohoku noson: Nihonmatsu-han Niita-mura* [Agricultural villages in northeastern Tokugawa Japan: The village of Niita in Nihonmatsu domain]. Tokyo: Dobunkan.
- Ochiai, E. (2001). Myth and reality of Asian traditional families: Living arrangement of the elderly in Tokugawa Japan. *Journal of Asian-Pacific Studies*, 9, 7–21.
- Ochiai, E. (2002). Kinsei josei hoko-nin nitotteno kon'in to shussan: Mutsu-no-kuni Asaka-gun Shimomoriya-mura Ninbetsu-Aratame-Cho no suryo-bunseki [Marriage and childbirth among female servants in the early modern period: A quantitative analysis of population registers of Shimomoriya]. *Josei Rekishi-Bunka-Kenkyujo-Kiyō* [Bulletin of Institute for Women's History and Culture], 10, 1–14.

- Ochiai, E. (Ed.) (2006a). *Tokugawa no kazoku to chiiki-sei* [Family and regionality of Tokugawa Japan]. Kyoto: Minerva-shobo.
- Ochiai, E. (2006b). Yurashia purojekuto no tassei: Rekishi-jinko-gaku to kazokushi [The achievement of the Eurasian Project: Historical demography and family history]. *Shakaigaku Kenkyu* [Journal of Social Science], 57(3-4), 57-80. Retrieved from <https://repository.dl.itc.u-tokyo.ac.jp/records/17193#.YXEVWBrP2bh>
- Ochiai, E. (2006c). Korei-sha no 'kodomu' tono dokyo: Tohoku-noson niokeru kaiso to kyoju-keitai [Elderlies living with 'children': Socioeconomic status and residential structure in rural northeastern villages]. In E. Ochiai (Ed.), *Tokugawa no raifu kosu* [The life course of Tokugawa Japan] (pp. 183-205). Kyoto: Minerva-shobo.
- Ochiai, E. (2009). Two types of stem household system in Japan: The *ie* in global perspective. In A. Fauve-Chamoux & E. Ochiai (Eds.), *The stem family in Eurasian perspective: Revisiting household societies, 17th-20th* (pp. 287-326). Bern: Peter Lang.
- Okada, A. (1998). Joto-gata koshu no tokucho: Mutsu-no-kuni Adachi-gun Niita-mura Ninbetsu-Aratame-Cho wo chushin toshite [Characteristics of the succession of heads: Based on Ninbetsu-Aratame-Cho of Niita, Adachi county, Mutsu province]. *Teikyo Shakaigaku* [Teikyo Sociology], 11, 109-135.
- Okada, A. (2000). Kinsei-nomin-shakai niokeru setai saikuru: Nihonmatu-han ni-kason no shiryo wo mochiite [The cycle of household structure in early modern peasant society]. *Shakaigaku Hyoron* [Japanese Sociological Review], 51(1), 136-152. doi: [10.4057/jsr.51.136](https://doi.org/10.4057/jsr.51.136)
- Okada, A. (2002). Kinsei-nomin-kazoku niokeru katoku no keisho to sono senryaku, Mutsu-no-kuni Asakagun Shimomoriya-mura Ninbetsu-Aratame-Cho wo chushin toshite [Succession and strategy among early modern peasants' family: Based on NAC of Shimomoriya, Asaka country, Mutsu province]. In A. Hayami (Ed.), *Kindai ikoki no kazoku to rekishi* [Family and history in the modern transition]. Kyoto: Minerva-shobo.
- Okada, A. (2006). *Kinsei Sonraku-shakai no ie to Setaikeisho: Kazoku-ruikei no Hendo to Kaiki* [Ie and household succession in early modern agricultural society: Variation and regression of family types (Doctoral dissertation)]. Tokyo: Chisen-shokan.
- Okada, A., & Kurosu, S. (1998). Succession and the death of the household head in early modern Japan: A case study of a northeastern village, 1720-1870. *Continuity and Change*, 13(1), 143-166. doi: [10.1017/S0268416098003099](https://doi.org/10.1017/S0268416098003099)
- Ono, Y. (1993). Bunka-kei no keisanki riyo II: Deta-nyuryoku no yuzaa intaafaisu, rekishi-jinko-gaku no baai [Computer utilization for humanities II: User interface for data entry, the case of historical demography]. *Nihon Kenkyu* [Bulletin of International Research Center for Japanese Studies], 8, 165-182.
- Saito, O. (1998). Two kinds of stem-family system? Traditional Japan and Europe compared. *Continuity and Change*, 13(1), 167-186. doi: [10.1017/S0268416098003087](https://doi.org/10.1017/S0268416098003087)
- Saito, O. (2000). Marriage, family labour and the stem family household: Traditional Japan in a comparative perspective. *Continuity and Change*, 15(1), 17-45. doi: [10.1017/S026841609900346X](https://doi.org/10.1017/S026841609900346X)
- Takahashi, M. (2005). *Zaigo-machi no rekishi-jinko-gaku: Kinsei ni okeru chiiki to chiho toshi no hatten* [The historical demography of Koriyama, a post town: The development of a local town within a community in early modern times] (Doctoral dissertation). Kyoto: Minerva-shobo.
- Takahashi, M. (2012). Zaigo-machi no kekkon to saikon [Marriage and remarriage in a local post town]. In S. Kurosu (Ed.), *Rekishi-jinko-gaku karamita kekkon, rikon, saikon* [Marriage, divorce and remarriage from the perspective of historical demography] (pp. 119-139). Kashiwa: Reitaku University Press.
- Takahashi, M. (2018, August). The use of horses in the Nihonmatsu domain, Tohoku district in Japan. Paper presented at the session "Subsistence, Sustainance, and Changing Living Spaces: Comparative Studies of Eurasian Economies from the 16th-20th centuries" at the World Economic History Congress, Boston, MA.
- Takahashi, M. (2019). The labour market and labour migration in small post towns in early modern Japan: The relationship between a town and its outlying villages in the northeastern domain of Nihonmatsu in the eighteenth to nineteenth centuries. In N. Okuda & T. Takai (Eds.), *Gender and family in Japan* (pp. 3-31). Springer.
- Takahashi, M., & Kurosu, S. (2020). Kinsei Nihon no jinko to kiko [Population and climate in early modern Japan]. In T. Nakatsuka, K. Kamatani, & K. Watanabe (Eds.), *Kiko hendo kara kinsei wo minaosu: Suryo, sisutemu, gijutsu* [Reviewing the early modern period from the perspective of climate change] (pp. 51-96). Kyoto: Rinsen-shoten.
- Tsuya, N. O., & Kurosu, S. (2002). The mortality effects of adult male death on women and children in agrarian household in early modern Japan. In R. Derosas & M. Michel Oris (Eds.), *When dad died: Individuals and families coping with family stress in past societies* (pp. 261-299). Bern: Peter Lang.

- Tsuya, N. O., & Kurosu, S. (2004). Mortality and household in two Ou villages, 1716–1870. In T. Bengtsson, C. Campbell, J. Z. Lee, et al. *Life under pressure: Mortality and living standards in Europe and Asia, 1700–1900* (pp. 253–292). Cambridge, MA: MIT Press.
- Tsuya, N. O., & Kurosu, S. (2005). Demographic responses to short-term economic stress in eighteenth and nineteenth-century rural Japan: Evidence from two northeastern villages. In R. Allen, T. Bengtsson, & M. Dribe (Eds.), *Living standards in the past: New perspectives on well-being in Asia and Europe* (pp. 427–460). Oxford: Oxford University Press.
- Tsuya, N. O., & Kurosu, S. (2010a). To die or to leave: Demographic responses to famines in rural northeastern Japan, 1716–1870. In S. Kurosu, T. Bengtsson, & C. Campbell (Eds.), *Demographic responses to economic and environmental crises* (pp. 79–106). Kashiwa: Reitaku University. Retrieved from <http://iussp.org/sites/default/files/AllArticles.pdf>
- Tsuya, N. O., & Kurosu, S. (2010b). Reproduction and family building strategies in early modern Japan: Evidence from two northeastern farming villages. *The History of the Family*, 15(4), 413–429. doi: [10.1016/j.hisfam.2010.05.004](https://doi.org/10.1016/j.hisfam.2010.05.004)
- Tsuya, N. O., & Kurosu, S. (2010c). Family, household, and reproduction in two northeastern Japanese villages, 1716–1870. In N. O. Tsuya, F. Wang, G. Alter, J. Z. Lee, et al. *Prudence and pressure: Reproduction and human agency in Europe and Asia, 1700–1900* (pp. 249–285). Cambridge, MA: MIT Press.
- Tsuya, N. O., & Kurosu, S. (2013). Social class and migration in two northeastern villages 1716–1870. *The History of the Family*, 18(4), 434–455. doi: [10.1080/1081602X.2013.815126](https://doi.org/10.1080/1081602X.2013.815126)
- Tsuya, N. O., & Kurosu, S. (2014). Economic and household factors of first marriage in two northeastern Japanese villages, 1716–1870. In C. Lundh, S. Kurosu, et al. *Similarity in difference: Marriage in Europe and Asia, 1700–1900* (pp. 349–391). Cambridge, MA: MIT Press.
- Tsuya, N. O., & Kurosu, S. (2017, June). Socioeconomic and family factors of first marriage: A comparative analysis of early modern and contemporary Japan. Presentation at the Population Association of Japan, Tohoku University, Japan.
- Tsuya, N. O., & Kurosu, S. (2019). Patterns and factors of first marriage: A comparative analysis of early modern and contemporary Japan. *Keio IES Discussion Paper Series* (No. DP2019–012). Tokyo: Keio University. Retrieved from <https://ies.keio.ac.jp/en/publications/11327/>
- Tsuya, N. O., & Kurosu, S. (2020, November). *Household socioeconomic status and mortality at different stages of life: Evidence from three northeastern Japanese villages, 1708–1870*. Presentation at the Population Association of Japan, Saitama Prefectural University, Japan.
- Tsuya, N. O., Wang, F., Alter, G., Lee, J. Z., et al. (2010). *Prudence and Pressure: Reproduction and human agency in Europe and Asia, 1700–1900*. Cambridge, MA: MIT Press.

HISTORICAL LIFE COURSE STUDIES
VOLUME 13 (2023), published 30-03-2023

The Demographic Database — History of Technical and Methodological Achievements

Pär Vikström

Maria Larsson

Elisabeth Engberg

Sören Edvinsson

Centre for Demographic and Ageing Research, Umeå University, Sweden

ABSTRACT

The Demographic Data Base (DDB) at the Centre for Demographic and Ageing Research (CEDAR) at Umeå University has since the 1970s been building longitudinal population databases and disseminating data for research. The databases were built to serve as national research infrastructures, useful for addressing an indefinite number of research questions within a broad range of scientific fields, and open to all academic researchers who wanted to use the data. A countless number of customized datasets have been prepared and distributed to researchers in Sweden and abroad and to date, the research has resulted in more than a thousand published scientific reports, books, and articles within a broad range of academic fields. This article will focus on the development of techniques and methods used to store and structure the data at DDB from the beginning in 1973 until today. This includes digitization methods, database design and methods for linkage. The different systems developed for implementing these methods are also described and to some extent, the hardware used.

Keywords: Database, Linkage, RDBMS, Digitization, Church registers

DOI article: <https://doi.org/10.51964/hlcs12163>

© 2023, Vikström, Larsson, Engberg, Edvinsson

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

Established in the infancy of historical demography, as Lionel Kesztenbaum in a recent article defines the period that ended in the early 1970s, the Demographic Data Base (DDB) at Umeå University has experienced remarkable technical and methodological development (Kesztenbaum, 2021). Manual excerpts, punch cards, and magnetic tapes are long gone, replaced by automated and semi-automated processes, custom-made software, and modern server technology (Edvinsson & Engberg, 2020). This article gives a brief description of the different technical approaches and methods used by the DDB to design, build, structure, and store large population databases since the start in 1973. In the first part, particular attention is given to the development of different structures to store data, and to digitization systems and hardware. The second part focuses on the development of different linkage methods and processes, from manual and semi-manual methods to fully automated processes.

Today, DDB owns and administers three main research databases: a) POPUM, with individual-level data from Swedish parishes in different areas, see Figure 1, covering the period 1680–1900; b) POPLINK, with similar data but covering a longer time span, until around 1950; and c) TABVERK, with aggregate statistics from all Swedish parishes for the period 1749–1859. POPUM and POPLINK are some of the most detailed historical databases in the world when it comes to the wealth of information per individual. The production database KBGRUNDS stores all digitized information from these registers, see Table 1 with an overview of the data as released on 2021-11-15. The latest version of POPUM and POPLINK, version 6.4.1, based on this release of KBGRUNDS, was released on 2021-12-16, see Table 2.

Already from the start, the databases at the DDB were built to serve as research infrastructures, useful for addressing an indefinite number of research questions within a broad range of scientific fields, and open to all academic researchers who want to use the data (Edvinsson & Engberg, 2020). The databases are based on the Swedish parish registers, which from 1680 until 1990 served as the system of official registration. Hence, the registers do not only include the majority population belonging to the Lutheran national state-church system, but also the part of the population belonging to other faiths and denominations. This means that the Swedish parish registers have a nearly unsurpassed coverage of the entire population. The records include births, marriages, and deaths, as well as family-based continuous registers covering the entire population, and their long time spans, from the late 17th century and forward, offer virtually unparalleled possibilities for longitudinal studies (Nilsdotter Jeub, 1993).

Although the technical and methodological changes have been immense, the principles guiding the basic demands of the longitudinal population databases have remained fairly unchanged (Johansson & Åkerman, 1973; Vikström, Edvinsson, & Brändström, 2006):

1. The database shall be *true to the source*. It must be possible to trace all records back to the original source for verification.
2. The database shall be *complete*; that is, all relevant information in the original source shall be included in the data collection.
3. The data collection shall be *open*, which means that the database shall be built in a way that allows the inclusion of new data, in time as well as in space.
4. The database shall be *coherent* and *consistent*: data entry shall be performed according to similar rules and principles, for maximum comparability and coordination.
5. Data entry, processing, and storage shall be performed in an *efficient* way.
6. All processing of data shall be *research-oriented*, allowing for micro-historic research as well as large-scale cohort studies.

Similar principles have later also been formulated by others, for example by Mandemakers and Dillon (2004) as best practice in the field of building longitudinal historical databases for research (Edvinsson & Engberg, 2020).

Figure 1 Geographical coverage of POPUM and POPLINK version 6.4.1

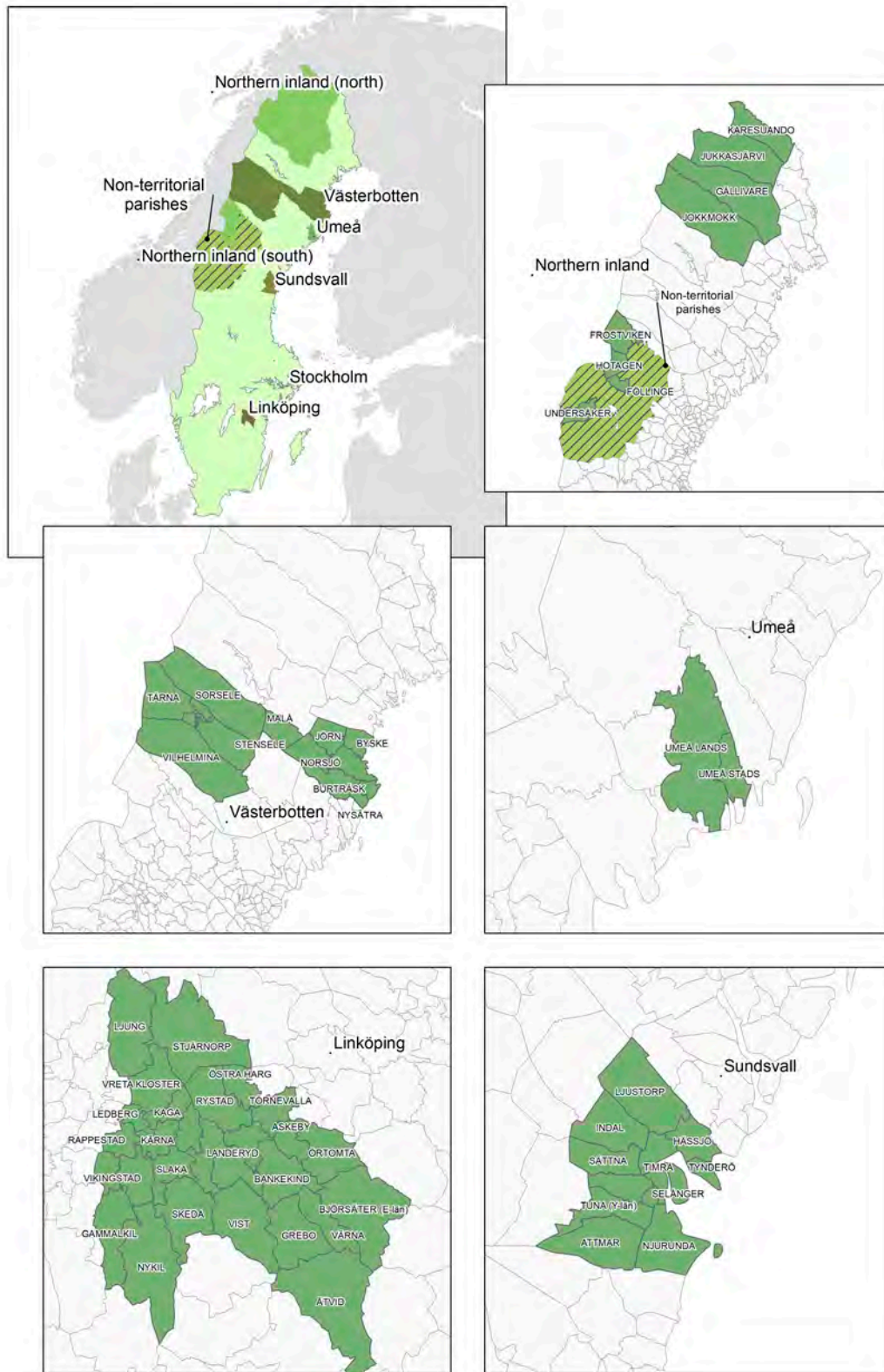


Table 1 *Number of source records in KBGRUNDS 2021-11-15*

Source (register)/ Region	Linköping	Northern inland	Sundsvall	Umeå	Västerbotten	TOTAL
Death and burial	167.621	30.898	50.235	20.452	150.793	419.999
Birth and baptism	229.745	54.856	105.623	40.149	342.288	772.661
Longitudinal parish	1.605.805	217.843	537.930	293.599	1.831.942	4.487.119
In-migration	202.147	7.048	83.358	67.254	161.383	521.190
Banns and marriage	35.243	13.438	25.453	12.505	76.928	163.567
Out-migration	225.284	5.461	72.232	58.061	186.516	547.554
TOTAL	2.465.845	329.544	874.831	492.020	2.749.850	6.912.090

Table 2 *Number of individuals in POPUM and POPLINK version6_4_1 (released 2021-12-16)*

Region	Linköping		Northern inland		Sundsvall	Umeå	Västerbotten		TOTAL	
Sex/Database	POPUM	POPUM	POPLINK	POPUM	POPLINK	POPUM	POPLINK	POPUM	POPLINK	
Unknown	5.618	1.033	54	189	9	1.141	1.328	7.981	1.391	
Male	325.339	46.073	6.243	88.673	56.142	114.240	198.574	574.325	260.959	
Female	334.024	45.063	6.543	83.021	57.777	111.313	191.924	573.421	256.244	
TOTAL	664.981	92.169	12.840	171.883	113.928	226.694	391.826	1.155.727	518.594	

2 DATABASE STRUCTURE AND DATA ENTRY SYSTEMS

2.1 DATA MANAGEMENT SYSTEMS BEFORE THE ERA OF RELATIONAL DATABASES

In 1973, when the data collection for the DDB-databases began, data entry was essentially a manual process. Between 1973 and 1982, data entry centers were established in six different locations in northern Sweden, funded by provisional contributions from the National Labour Market Board. At the height of operations, more than 100 data entry assistants worked at these centers. Information from the sources was transcribed by hand into forms on printed cards, one card for each entry. These paper cards were then manually linked together by ordering them into bundles, each bundle describing one individual, and the information was digitized using punch cards (Edvinsson & Engberg, 2020). Two technical hubs for this kind of work were set up, one in Umeå and one in Haparanda, where the very first data entry centre was established. After a couple of years the punch cards were abandoned and a commercial digitization system, also called "KEY to DISK", was introduced, and the bundles of cards were digitized on small terminals. This made the process of digitization more efficient. The first system of this kind was a CMC 12 (Communication Machinery Company) upgraded after five years to a CMC 5400.

The Swedish parish registers are ordered into an intricate system, making up something that almost resembles a kind of non-digitalized relational database structure. The base in the system is the catechetical register, a household-based longitudinal parish register which is a distinctive feature of the Swedish registers, providing basic demographic information about the entire population in a parish. Like in a census, families were kept together on the same page. When a child was born, a parent died, or a widow remarried, it was not only noted in the separate event registers, but also in the longitudinal register. Record linkage is facilitated by clever links between said registers. The longitudinal registers include references to the volume and page, where first-hand information about a particular birth, marriage, or death can be found. The event registers are linked in a similar way to the longitudinal registers, creating a comprehensive system of information whereby individuals can be followed over their entire life spans. With this kind of double bookkeeping, it is also possible to reconstruct missing or

lost information. When a longitudinal register volume was completed after five to ten years a new one was established, into which the minister transferred current information from the old volume, of course with valuable links between the old and new registers. For obvious reasons, the longitudinal registers are extremely valuable for the creation of life-course data, making it possible to follow individuals and families over their entire life spans and over generations as long as they remain within the parish borders (Edvinsson & Engberg, 2020).

This comprehensive structure of the sources was at the beginning used as a template to build a database structure consisting of flat files, called Individ. These files served well to store the information, but the flat format was not optimal for data extraction, both in the case of defining a cohort for research but also in extracting the variables required. The Individ-files also had other limitations. Due to the digitization systems used during the first years, before relational databases were introduced, the records had a maximum length of 256 characters, which usually resulted in one record from the source becoming two records in the file. To solve this problem, the Individ-files were converted into a 518-character format, called INDIVSQ, which covered the full record from the source. This conversion was made by the DDB, after the information had been transferred to hardware at the computer centre at Umeå University, UMDAC. Much of the programming and other information processing at the DDB was, during the end of the 1970's and the beginning of the 1980's, performed by using a terminal connection to UMDAC and portable printing terminals, Texas Silent.

Despite the limitations in database structure, a large amount of data was collected and digitized during these early years. Records were transcribed, linked manually and stored in a database management system. The first parish to be digitized was Tuna, a small parish in the industrialized Sundsvall area, which at that time was already an object of interest among historians and social scientists as being one of the fastest growing industrial regions in Sweden (Brändström, 2009). The work continued with records from six other parishes in different Swedish regions, Locknevi, Gullholmen, Trosa, Fleninge, Nedertorneå and Svinnegarn, selected on scientific merits expressed by researchers. Nedertorneå, was for example chosen because of its high level of infant mortality (Brändström, 1984). Together, these seven parishes formed the beginning of the database that later would be named POPUM. The first large region included in POPUM was the industrialized Sundsvall area, a previously agricultural district that in the 19th century became the heart of the sawmill industry in Europe and a center for the Swedish labor movement. In the early 1980s, the Skellefteå area in the north was selected for a large project in genetic epidemiology (Edvinsson & Engberg, 2020). By then, small Luxor ABC802-computers were used for digitization. The diskettes were continuously transferred by mail from the data entry centers to the technical hub in Umeå, copied to the computers at UMDAC and later converted to INDIVSQ-format.

2.2 RELATIONAL DATABASE MANAGEMENT SYSTEMS

But soon, a new and influential general theory of data management would change the situation and solve the problems with the flat files. In 1970 Dr E. F. Codd presented his seminal work "A Relational Model of Data for Large Shared Data Banks" (Codd, 1970), where he presented a theoretical model for relational databases and relational database management systems that would have a large impact upon database modelling and significantly facilitate the development of large databases, including the DDB.

In 1983 DDB decided to start converting all data into a relational database, using the model that E. F. Codd had presented 13 years earlier. The first relational database management system (RDBMS) used by DDB was IBM SQL/DS. With this new system it was possible to use the SQL standard language to select and retrieve data from the database, and to join information from different tables. Making simple retrievals was easy to accomplish.

However, the transition from the old to the new database management system did not take place without difficulties, and it turned out to be a more time-consuming enterprise than originally planned. The first preparations for the conversion were made in 1983, by specifying demands on the new database structure and the practical work started in 1985. Almost all hardware had to be installed at the same time. It would have been better if one of the systems had been installed as a pilot system, making it easier to identify problems and how to solve these, before the second system was installed.

The hardware consisted of a minicomputer system from IBM called 4361 VM/CMS, one in Umeå and one in Haparanda. A somewhat special situation occurred when installing the hardware in Haparanda. A window, and a part of the wall of the building had to be removed, to be able to lift in the minicomputer, as this quite huge machine actually was called.

In this novel technical environment, digitization, programming, and database management were performed from terminals connected directly to the two minicomputers. New software had to be developed for the conversion of INDIVSQ-files to a relational database and a new data entry system also had to be designed and developed to interact with the technical environment. At the same time the staff at DDB also had to be trained to be able to adapt to a completely new way of data processing.

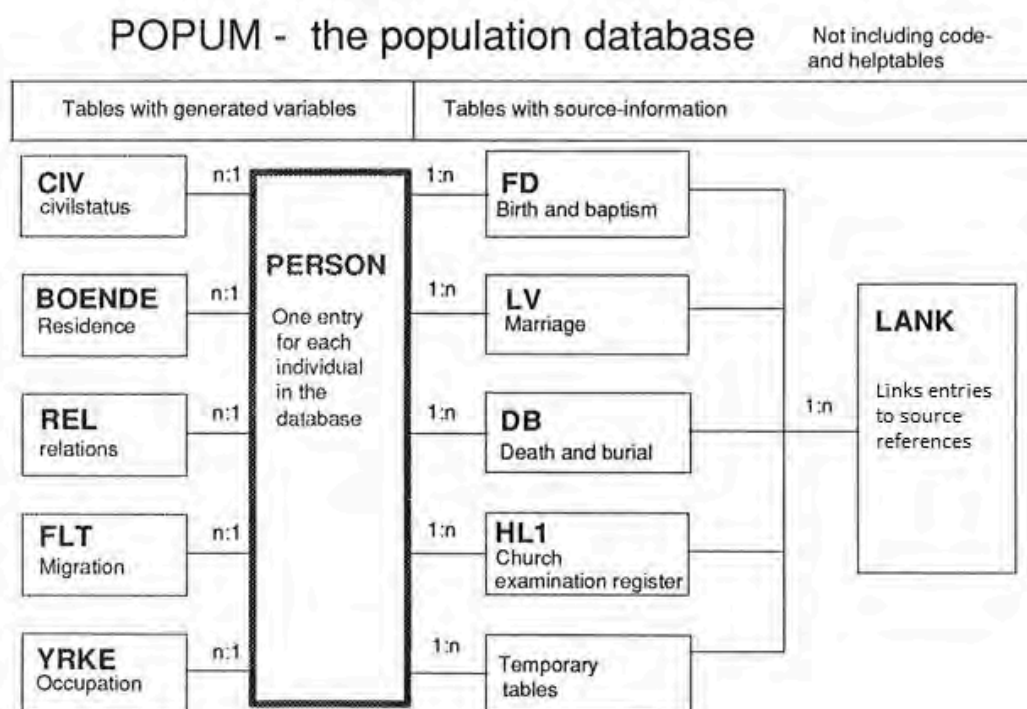
This whole situation caused several problems. One was that the pace of digitization slowed down. Another, even more crucial problem, was that the research using DDB-data also slowed down. For a period of almost three years very few retrievals for researchers were made.

After more than two years, in 1987, the very first version, a beta version, of the relational database was ready for use. It did not have a name until version 2 of the database was published in 1991. Then the database was given the name POPUM, made up from two central concepts, Population and Umeå. An overview of this early relational database is presented in Figure 2.

So, why did the conversion end up taking three years instead of the nine months as originally planned? The answer is simple. It was pioneering work. No one at DDB had made this kind of transition before. However, knowledge and competence accumulated during this process has made later transitions easier.

Although the new relational database management system significantly improved the data production process, there were still technical issues that caused problems and required attention. The most problematic issue was perhaps limitations in disk space. Disks were extremely expensive, and the two minicomputers did not have enough capacity to handle the required amount of data. A lot of time was spent finding solutions to work with large amounts of data from big parishes. The solution was to store data from one part of the parish on magnetic tapes, while dealing with the other part on disk. In 1994 the issue of disk space became less problematic, when the old IBM 4361 system gave way for new IBM RS6000-servers with the IBM dialect of UNIX, AIX. All terminals were substituted with personal computers connected in a network. At the same time, the relational database management system was changed from SQL/DS to Ingres, mostly because SQL/DS did not work with AIX. A new data entry system, RODE, was also developed and used on the personal computers. Files from RODE were transferred from the personal computers over the network to a production database, POPHAP, for subsequent processing of the data, involving coding, standardization, and linkage, before finally being included in POPUM. POPHAP was used until 2005, in the end mostly for linkage purposes.

Figure 2 POPUM version 2



2.3 A NEW MILLENNIUM

At the end of the previous millennium the RS6000-servers were getting old and had to be replaced. Replacing them with new RS6000 would have been too expensive, so in 1998 it was decided that they should be replaced with INTEL-based servers from DELL and HP. INGRES was replaced with IBM DB2 in 1999. Around the same time, a new digitization system, REGINA, was developed, primarily for adaption to the new database management system but also to avoid problems with possible millennium bugs.

Over the years it had become evident that not all the requirements on the database structure set in 1983 had been met. This, in combination with a normalization that needed some adjustments and new demands caused by methodological developments in research, made it necessary to conduct a larger review of the database structure. Another issue that had become more and more disturbing over the years, was the use of multiple digitization systems for entering data. The data, and in particular the coding schemes from the different systems were not always harmonized, which caused a lot of extra programming when data was extracted for researchers. These issues were finally solved with version 3 of POPUM, in 2006.

Along with version 3 of POPUM a new database was added to improve the normalization and avoid inconsistency in data, the database KBGRUND. This new database constitutes the basic database at DDB, containing all information that has been digitized, and into which all new data is stored and post processed. All information in KBGRUND is traceable back to the original source, not only to the record but also to the individuals mentioned in the source record, i.e., in the birth register we have, not only the child, but also its mother and father. KBGRUND is a strictly normalized database. POPUM version 3, on the other hand, is a user database containing specific post-processed tables describing certain events and states, like migration and residence, as well as all tables from KBGRUND. POPUM is mainly used for retrieval of subsets of the data for research and therefore a certain degree of denormalization is allowed. But to avoid inconsistencies, the database is only available as read-only. Version 3 of KBGRUND was released at the same time as REGINA in 2000 and version 3 of POPUM was released in 2006.

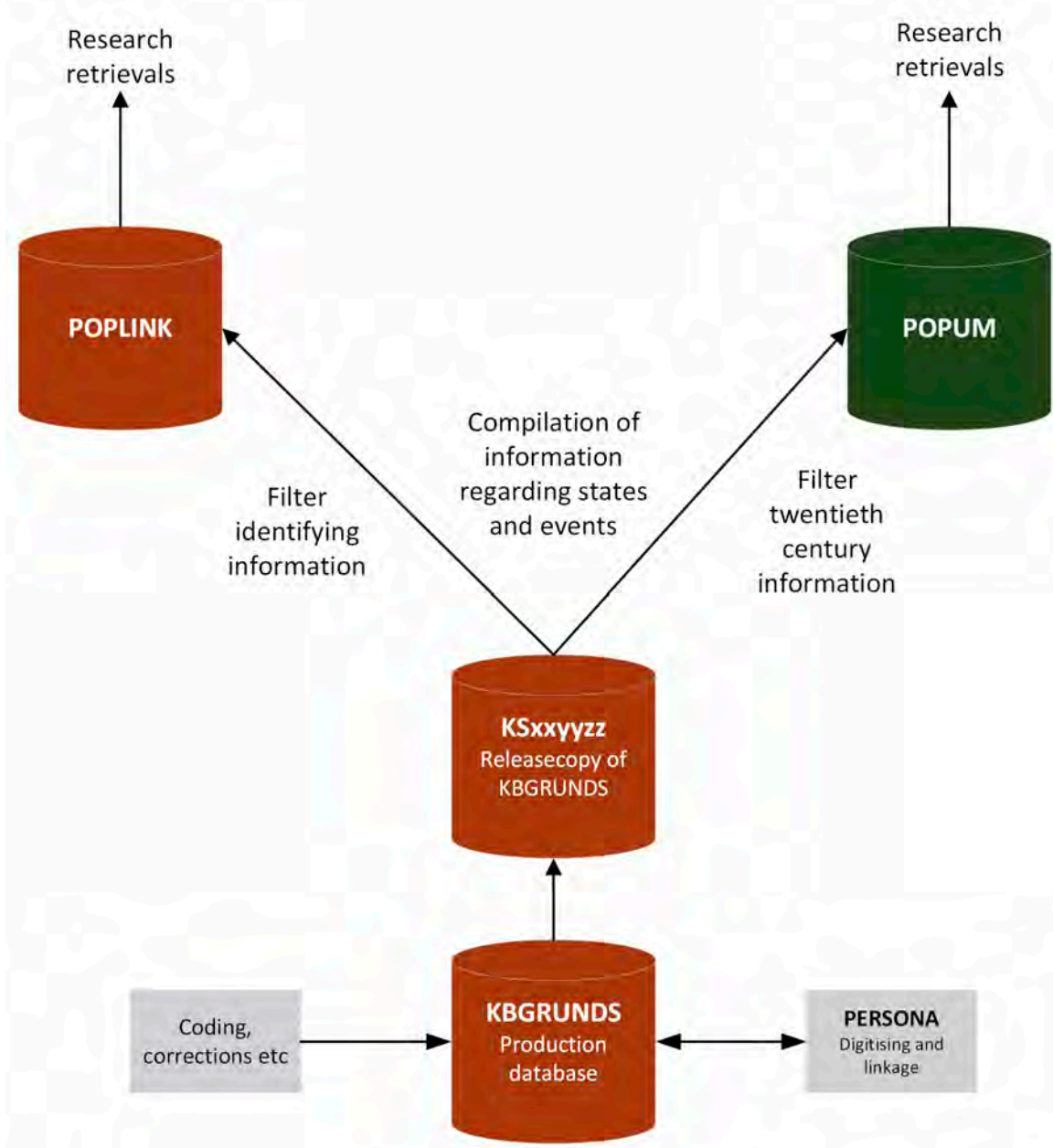
2.4 THE POPLINK DATABASE — INCREASED DEMANDS ON INFORMATION SECURITY

In 2008 the scope of the data collection at the DDB was extended. Before that, focus had been on data from the 18th and 19th centuries and data collection had usually ended around the years 1900–1910. In a new project, data from the 20th century from two regions in northern Sweden was digitized, allowing researchers to take advantage of the long time spans in the Swedish sources. The objective was to create a new asset for micro-level analysis of the processes that transformed society, through the demographic transition and beyond, by bridging the present gap between historical population databases and modern registers (Westberg, Engberg, & Edvinsson, 2016).

Handling individual-level data stretching as far as 1960 raises several ethical issues, adding a new level of demands on the information- and IT-security. Already before the implementation of the European General Data Protection Regulation¹ (GDPR) in 2018, Sweden had a strict legal framework concerning the protection of privacy, with implications for the data production process as well as for the release of data for research. The extended scope of data collection, including information that constitutes personal data according to the GDPR, meant that some changes had to be made regarding the storage of the databases. The production process including the production database KBGRUND (from now on KBGRUNDS) was moved to a secure encrypted network. A new population database, POPLINK, with the same structure as POPUM, was created in the secure network to handle retrievals containing 20th century data. The secure network safeguards data with a double layer of physical security and access authentication. The population database POPUM, containing no information about living persons, and thus no personal data, is still stored in the "open" network to make this information easier to retrieve (Westberg, Engberg, & Edvinsson, 2016). The information in the production database KBGRUNDS is filtered in different ways to the population databases POPUM and POPLINK, as shown in Figure 3. The red databases are in the secure encrypted network, and the green database in the "open" network.

1 <https://gdpr.eu/>

Figure 3 *Compilation of POPUM and POPLINK*



The first version of POPLINK was built by adding new data for the period 1900–c1960 to parishes already existing in the POPUM database at the DDB to reduce the time between data entry and release of the data. The linkage between old and new data worked almost seamlessly and POPLINK has since then been a valuable resource at the DDB and for the research community. Data is continuously added to the database as new parishes are digitized and linked. New versions of POPUM and POPLINK are released about once a year and as part of this process new decisions are made regarding which information must be stored in the secure encrypted environment.

2.5 THE NEW DIGITIZATION SYSTEM PERSONA AND VERSION 6 OF KBGRUNDS

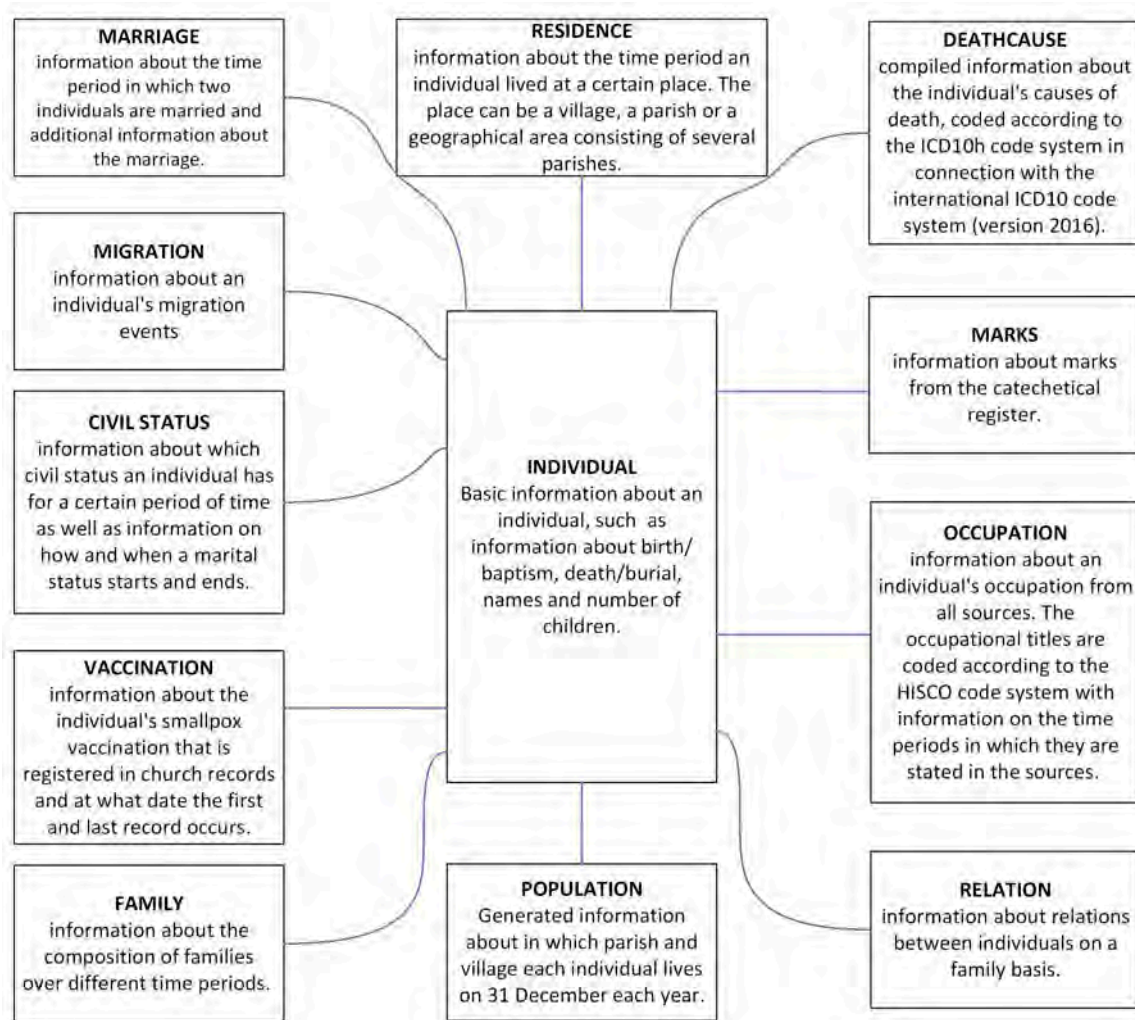
In 2012 the DDB was granted funding from the Swedish Research Council, to develop a new digitization system. The old system Regina had been used for 12 years, and during that period a number of revisions had been made to the software. Moreover, the software and programming languages used for developing REGINA were at the time becoming outdated, so it was due time to develop a new system instead of making new revisions.

But this time the aim was higher than just serving the needs of the DDB. The idea was to develop a web-based tool for database building that could serve the entire research community, not only the

DDB, increasing the interoperability of population databases in Sweden. During the project period, the software was named PERSONA. However, building a tool intended for a more general use called for a higher level of harmonization and standardization, in database modelling as well as in coding of variables, using national and international standards. Version 6 of KBGRUNDS was designed to be a strictly normalized database adapted to the new digitization system PERSONA.

The first version of PERSONA was put into operation in 2016 and has since then been used for all digitization and linkage of parish registers at DDB. The system is at present also used by the Swedish National Archives and by Gothenburg University (the Gothenburg Population Panel). All the software needed to produce a new version of POPUM and POPLINK, version 6, had to be redesigned to fit version 6 of KBGRUNDS. A major change was also made in the structure and contents of the new POPUM and POPLINK, as shown in Figure 4. As the production database is strictly normalized it is difficult to retrieve and analyze data. To facilitate the work with data retrievals, make it easier for users to retrieve data, the new versions of POPUM and POPLINK only include tables with compiled information describing certain events and states of an individual. It took some time to make these changes, but the result was worth the wait. Most research retrievals can now be made much faster than with earlier versions of the databases.

Figure 4 *POPUM and POPLINK version 6.4*



3 DEVELOPMENT OF THE LINKAGE PROCESS

From the infancy of the DDB it was evident that linkage of the digitized records would be required to create an efficient research database for life-course studies. Numerous records from different sources had to be ordered into a continuous chronology, constructing biographies, which in the ideal case cover the entire life span of an individual from the cradle to the grave. The presence of kinship links and family relationships has also significantly increased the scientific value of and usefulness of the data.

3.1 TOWARDS A SEMI-AUTOMATED PROCESS

As already mentioned, the very first form of record linkage was to order and group the paper cards with transcribed entries into bundles, one bundle representing one individual (Edvinsson & Engberg, 2020). Linkage of family relationships was also done manually, by way of a form of traditional family reconstitution. Along with the advances of computer technology, these manual procedures were replaced by semi-automated and computer-aided linkage systems.

In the late 1990s the first linkage software, ManLank (Manual Linkage), was developed. It was a computer-aided system that supported a manual linkage of individual records and of family relationships within a parish. At that time only two specially appointed personnel were allowed to work with linkage. They had no time pressure to finalize the linkage of a parish and used a lot of time scrolling backwards and forward in the sources to find matches.

Between 1999 and 2002, a semi-automated linkage process, using a combination of automated and manual methods of record linkage was developed and implemented. The most important improvement was the development of new software for automated linkage, CoreLink (Computerized Record Linkage), described in detail in section 3.3.1. Linkage was now performed by all data entry assistants, making linkage a fully integrated part of the data production process. In the first decade of the new millennium the linkage process consisted of three steps: a) an automated record linkage with CoreLink, b) a computer-aided record linkage with an updated version of ManLank, and c) a manual linkage of relationships.

Although an automated linkage phase, a), was introduced, the computer assisted manual linkage step, b), was maintained. The aim was twofold: 1) to perform linkage of records that could not be linked by the software, and, at the beginning, also 2) to validate the result from the automated record linkage process. In all kinds of linkage, secure links has always been a main priority for the DDB, far more important than a high linkage rate. For the manual linkage, all available information in the sources could be used to validate and link information that could not be linked by the software. For instance, if information about an individual was scarce, records with information about parents and partners could be tracked and used to make a secure linkage. At the beginning, all links made by CoreLink were manually scrutinized, but after some years the manual process was changed into a residual linkage, only focusing on linked records that were flagged by the system as incomplete or inconsistent and needing manual attention. The main reason for changing the process was that scrutinizing every link was very time-consuming, and, as very few changes were made, it had almost no impact on the result. Examples of links still being scrutinized, as they are flagged, are for instance individual records where the automated system has detected gaps in life-biographies, which are not associated with migration. Another common scenario is when a date for a birth, marriage or a death is noted in the longitudinal parish register, without a matching record in the event registers linked to the biography. The last step in the linkage process, c), was a linkage of relations, that is, linking parents to children and spouses to each other. This linkage step was performed manually with computer assistance. All relation links were stored in a specific relation table in the database, a table which still is instrumental for constructing genealogies over several generations.

3.2 FURTHER LINKAGE STEPS AND INCREASED AUTOMATIZATION

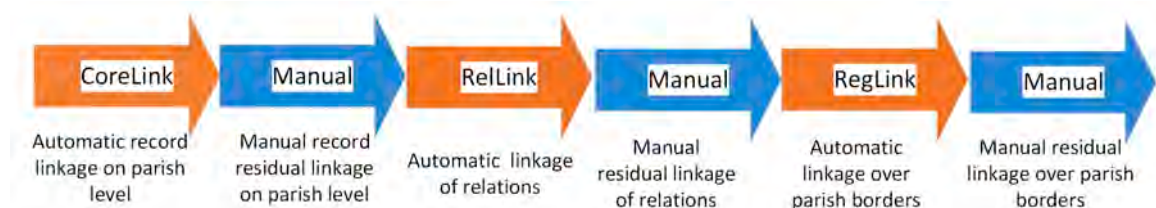
Since the digitization and linkage of parish records mostly have been performed register by register for one parish at a time, a high linkage rate on the parish level does not necessarily imply complete life-course data at the individual level. It is particularly migration back and forth over parish borders that causes problems. The registers do not cover events that took place in other parishes before or after migration, and therefore most migrants have incomplete life course data or apparent gaps in their life-biographies. Another problem is caused by administrative changes in parish structure, resulting

in that a distinct part of the population disappears from the sources for other reasons than migration at a fixed point in time. A problem that became more and more evident as the database expanded was duplets, caused by short-distance migration, that is, people moving between adjacent parishes, sometimes several times. Some of these migrants ended up having two or more different identities in the database: one in their parish of origin and other ones in parishes they later belonged to. A number of steps have been taken to solve these issues.

The problem with information loss due to short distance migration and administrative changes of parish borders was solved by introducing an additional linkage step, linking larger areas consisting of adjacent parishes. Accounting for short distance migration significantly increases the number of complete life-courses in the data. At first, this linkage step was performed manually using SirLink (Simple Regional Linkage), an in-house developed software. However, it ended up being a time-consuming enterprise and soon it became obvious that assistance from an automated linkage software was needed. Since CoreLink, was focused on linking records within a parish and this kind of linkage needed a completely different approach, it was decided to develop new software. In 2010, new software called RegLink (Region Linkage) was added to the linkage process allowing for automatic linkage of individuals across parishes.

In 2014, the most recent addition to the linkage process was made when RelLink (Relation Linkage) an automatic software for linking relations, was developed. The linkage process, as shown in Figure 5, was thereby completed with both automated and manual steps for all three types of linkage. With these new automatic steps, the linkage process time was significantly reduced. In 2016 finally, all manual steps of linkage were included in the same software, the production system PERSONA, described above.

Figure 5 *Current Linkage Process*



3.3 SOFTWARE FOR AUTOMATED LINKAGE – TECHNICAL ASPECTS

The development of automated linkage software at DDB described in 3.1-3.2 began with an evaluation of different linkage approaches. A probabilistic approach, as used when linking American censuses (see for example Abramitzky, Boustan, Eriksson, Feigenbaum, & Pérez, 2021), was compared to a rule-based approach. Different probabilistic methods were also tested.

The data from Swedish church records is of high quality and has, as described in section 2.1, the advantage of the records being "pre-linked" by the minister who kept the books. The longitudinal parish registers have references to information in the event registers, and the event registers include distinct references to the longitudinal parish registers. The latter also include individual level references to previous and subsequent registers, making it possible to trace individuals between the registers, back and forth in time (Nilsdotter Jeub, 1993, p. 65). Therefore, it was decided that a rule-based approach where the references in the sources between registers could be used was to be preferred. The rule-based approach resulted in a high linkage rate with a very small amount faulty links.

3.3.1 CORELINK — AUTOMATED RECORD LINKAGE WITHIN A PARISH

In 2002 the first version of the software CoreLink for automated linkage was finalized and implemented in the production process. In this first version, four out of the five principal sources in the Swedish parish records were linked together, namely the longitudinal parish records, birth, marriage, and death registers. The migration registers were left out as they were considered too difficult to handle automatically, due to their frequently poor information about family members. Often, the registers included only the name of the head of the migrating family, or group of migrants, along with information about the number of men and women that were moving out or in together. The solution became to manually

add a record containing the same information to all migrants in the family group. In the second version of Core Link implemented in 2009, migration records were included in the automated linkage process. This version linked only the head of the migration group not the other individuals, because only the head of the group had enough information to make automated linkage possible. In the third version of the software, released in 2012, the unique national registration number, introduced in 1947–48 was included as a linkage variable, something that significantly improved the automated linkage of 20th century records.

As stated before, the linkage is rule based, and the primary aim of CoreLink is to use well defined algorithms and rules to decide if a link can be established or not. Matching is done between pairs of records or between groups of records. Exact matches and clear miss-matches are processed automatically using well-trying algorithms for searching and matching. The software gives all individuals in the data unique identity numbers. If there are records impossible to link to any other record, a new individual will be "created" with only one record.

The core of the Swedish parish registers are the longitudinal registers, with their comprehensive system of references between individual records and sources. They are therefore used as the main source for linkage. During the development of CoreLink, different sequences of linking event registers to the longitudinal registers were tested, to determine which sequence resulted in the best links. This resulted in that the following sequence was implemented: a) linking birth and baptism records to the longitudinal parish records; b) matching and linking together all individual records in the longitudinal registers; c) matching and linking migration records; d) matching and linking banns and marriage records, and finally, e) matching and linking death and burial records.

Since a person in a population often cannot uniquely be identified by his or her name — naming practices showed very little variation before 1900 — it was necessary to include other unique variables in the linkage. In the chosen model, sex was used as a blocking variable together with either year of birth or year of an event, such as year of marriage. The other variables used in different combinations were date of birth, parish of birth, date of event (death, marriage, migration) and page references made by the minister. Names are compared, using Levenshtein distance (<https://www.cuelogic.com/blog/the-levenshtein-algorithm>) and standardized names.² Since most common first names in Sweden are short, working with standardized names has been important to reach the 75% correspondence between the strings, that according to the definition is required for a match. A good example is the name PER, which also can be spelled PÄR or PEHR. Since in both cases the difference between the name variants is one letter out of three, the correspondence only reaches 66%, a measure that is considered too low to constitute a match. This is the reason why name standardization considerably improves the linkage rate

Documenting how a link has been made is important and therefore every decision made by the software was entered into a log table stating which rule a certain record was linked by. After the automated linkage, a special routine is executed to identify problems in the links such as illogical start or end dates, overlapping records, illogical gaps of time between records, the occurrence of more than one birth parish, errors in references between sources, etcetera. This information is used during the manual validation of the linkage, identifying possible faulty links.

3.3.2 RELINK — AUTOMATED LINKAGE OF RELATIONS

The Swedish parish registers offer ample information about relationships. In the longitudinal parish registers families were registered together, and the type of relation between the family members is normally specified for all individuals on each page. Parents are named in the registers of birth and death. Names of spouses and sometimes also parents are usually found in both marriage and death registers. As most individuals have multiple records, there is a need to create unique relation links. In 2014 the software RelLink for automated linkage of relations between parents and children and between spouses was developed and implemented. The software uses several different rules for each kind of relation. For example, information from birth records is used to determine whether a relation between a child and a parent is of biological nature. Information from marriage records and longitudinal parish records is used to define status for the relation between spouses: engaged, betrothed or married.

2 The Levenshtein distance is a metric for measuring the difference between two sequences.

3.3.3 REGLINK — AUTOMATED LINKAGE ACROSS PARISHES

The RegLink software matches and links individuals in different parishes, into one unique identity, making it possible to follow individuals moving from one parish to another within the same area in the data.

RegLink was constructed in a similar fashion as CoreLink, but includes a new, separate group of rules for linking individuals in different parishes. The rule-based linkage in RegLink is based on a number of identified scenarios. The first scenario uses the events, birth, marriage and death, from three different aspects.

1. If an individual has a record in a source in parish A, indicating that a birth, a marriage or a death took place in parish B, the software tries to find this individual in parish B, as noted in the register.
2. If an individual has information about birth, marriage or death in the longitudinal parish register where he/she is registered, without a matching record in the birth-, marriage-, or death register in the same parish, the software tries to find these records in other parishes.
3. If an individual in a birth, marriage or death record in parish A, is mentioned to reside in parish B, the software tries to locate this individual in a register from parish B.

The second scenario uses information about migration from migration registers and longitudinal parish registers. Links are made when an individual is leaving one parish and entering another parish in the same year, and with the same co-movers. When using this scenario surname is an important variable. Women, who frequently migrated in association with marriage and thus also might have changed their surnames, must be handled in a special way. If the date of marriage is close to the time of the migration, both the woman's maiden name and her new husband's surname are used as linking variables.

The third scenario covers administrative migration, that is, when an individual is transferred to another parish due to an administrative change, such as the detachments of a new parish or changed parish borders. Here, linkage requires that the individual is recorded at the same place of residence and with the same relatives in the parish register, both before and after the administrative migration, and that the year of the administrative migration is equivalent to the start year of the new parish.

A particular challenge has been handling individuals with periods of absence. One of the most difficult groups to link are single men and women moving out of the region covered by the database and later returning to a different parish than the one they once left. Trying to solve this problem, a set of rules looking for the absence of records of residence in the life-biography of an individual were added, aiming to find matching individuals for at least a part of that absence. Other variables used in these cases are date of birth, names, parish of birth, occupation and relatives. The linkage of individuals in this group has been improved by the automated linkage, but a small under-linkage remains.

4 CONCLUSIONS

During the 50 years that DDB has existed, the purpose of the infrastructure has remained the same: building high-quality longitudinal population databases for research, developing effective methods for database construction, and disseminating data for research. From the 1970s and until today, technological and methodological advancements and innovations within this field have been immense. Manual excerpts and rudimentary forms of linkage have been replaced by comprehensive digitization systems with processes, at the same time reducing the process time and improving the quality of the data. Early database models with limitations have given way for advanced and flexible database management systems, increasing data security and consistency and facilitating long-term management and data retrievals. A certain amount of manual validation of the results of the automated processes has however been kept ensuring the quality of data. Even though the digitization systems have improved, the actual interpretation of the text is still manual. Looking forward, handwritten text recognition (HTR) techniques stand out as an interesting development of the technical infrastructure, following the example of the BALSAC in Canada (Vézina & Bournival, 2020). Another important improvement is the national partnership within the SwedPop infrastructure, that besides its large value for future comparative research also has brought about valuable methodological collaborations with other Swedish databases, for example SEDD at Lund University.

REFERENCES

- Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., & Pérez, S. (2021). Automated linking of historical data. *Journal of Economic Literature*, 59(3), 865–918. doi: [10.1257/jel.20201599](https://doi.org/10.1257/jel.20201599)
- Brändström, A. (1984). "De kärlekslösa mödrarna": Nedgången i spädbarnsdödlighet i Sverige under 1800-talet, med särskild hänsyn till Nedertorneå (Doctoral dissertation). Umeå University.
- Brändström, A. (2009). Demografiska databasen och historisk demografi i Umeå — "En allvarlig felinvestering"? In R. Jacobsson (Ed.), *Thule: Kungliga Skytteanska Samfundets årsbok 2009* (pp. 211–222). Umeå: Kungliga Skytteanska Samfundet.
- Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, 13(6), 377–387. doi: [10.1145/362384.362685](https://doi.org/10.1145/362384.362685)
- Edvinsson, S. (2000). The Demographic Data Base at Umeå University: A resource for historical studies. In P. Kelly Hall, R. McCaa, R., & G. Thorvaldsen (Eds.), *Handbook of international historical microdata for population research* (pp.231–248). Minneapolis: Minnesota Population Center. Retrieved from https://international.ipums.org/international/microdata_handbook.shtml
- Edvinsson, S., & Engberg, E. (2020). A database for the future. Major contributions from 47 years of database development and research at the Demographic Data Base. *Historical Life Course Studies*, 9, 173–196. doi: [10.51964/hlcs9305](https://doi.org/10.51964/hlcs9305)
- Johansson, E., & Åkerman, S. (1973). "Faktaunderlag för forskning. Planering av en demografisk databas". *Historisk tidskrift*, 3, 406–414.
- Karlsson, T., & Lundh, C. (2015). *The Gothenburg Population Panel 1915–1943: GOPP version 6.0*. (Papers in Economic History No. 18). Lund University Publications. Available from <https://lup.lub.lu.se/search/publication/7870561>
- Kesztenbaum, L. (2021). Strength in numbers. A short note on the past, present and future of large historical databases. *Historical Life Course Studies*, 10, 5–8. doi: [10.51964/hlcs9557](https://doi.org/10.51964/hlcs9557)
- Lund, R. (2017). *Regler för konvertering av KBGRUNDS5 till KBGRUNDS6*. Centre for Demographic and Ageing Research, Umeå Universitet.
- Mandemakers, K., & Dillon, L. (2004). Best practices with large databases on historical populations. *Historical Methods*, 37(1), 34–38. doi: [10.3200/HMTS.37.1.34-38](https://doi.org/10.3200/HMTS.37.1.34-38)
- Nilsdotter Jeub, U. (1993). *Parish records: 19th century ecclesiastical registers*. Information from the Demographic Data Base. Umeå: Umeå University, Demographic Data Base.
- Vézina, H., & Bournival, J.-S. (2020). An overview of the BALSAC population database. Past developments, current state and future prospects. *Historical Life Course Studies*, 9, 114–129. doi: [10.51964/hlcs9299](https://doi.org/10.51964/hlcs9299)
- Vikström, P., Edvinsson, S., & Brändström, A. (2006). Longitudinal databases-sources for analyzing the life-course: Characteristics, difficulties and possibilities. *History and Computing*, 14(1–2), 109–128.
- Westberg, A, Engberg, E., & Edvinsson, S. (2016). A unique source for innovative longitudinal research: The POPLINK database. *Historical Life Course Studies*, 3, 20–31. doi: [10.51964/hlcs9351](https://doi.org/10.51964/hlcs9351)

HISTORICAL LIFE COURSE STUDIES
VOLUME 9 (2020), published 12-11-2020

The Scanian Economic-Demographic Database (SEDD)

Martin Dribe
Lund University, Sweden

Luciana Quaranta
Lund University, Sweden

ABSTRACT

The Scanian Economic-Demographic Database (SEDD) is a high-quality longitudinal data resource spanning the period 1646–1967. It covers all individuals born in or migrated to the city of Landskrona and five rural parishes in western Scania in southern Sweden. The entire population present in the area is fully covered after 1813. At the individual level, SEDD combines various demographic and socioeconomic records, including causes of death, place of birth and geographic data on the place of residence within a parish. At the family level, the data contain a combination of demographic records and information on occupation, landholding and income. The data for 1813–1967 was structured in the model of the Intermediate Data Structure (IDS). In addition to storing source data in the SEDD IDS tables, a wide range of individual- and context-level variables were constructed, which means that most types of analyses using SEDD can be conducted without the need of further elaboration of the data. This article discusses the source material, linkage methods, and structure of the database.

Keywords: Sweden, Historical demography, Family reconstitution, Population registers, Longitudinal data, Life courses, Intermediate Data Structure

DOI article: <https://doi.org/10.51964/hlcs9302>

© 2020, Dribe, Quaranta

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

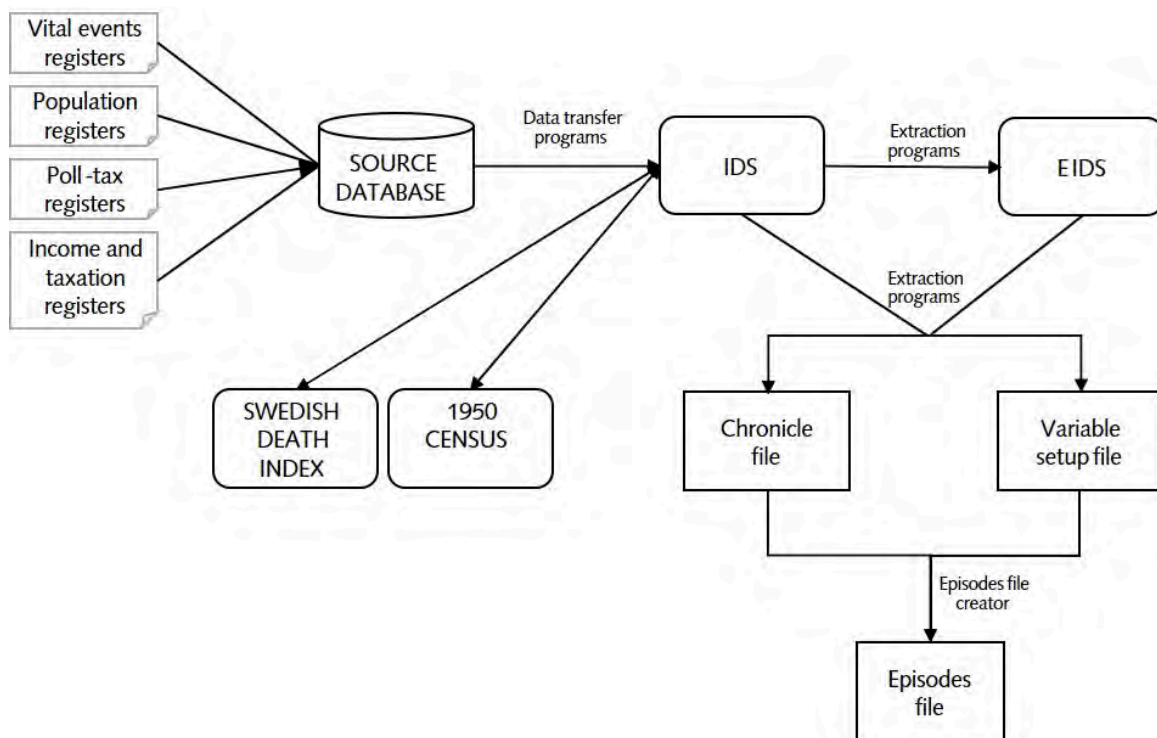
The Scanian Economic-Demographic Database (SEDD) is a high-quality longitudinal data resource spanning the period 1646–1967. It covers all individuals born in or migrated to the city of Landskrona and five rural parishes in western Scania in southern Sweden. Within a specific research program at the Center for Economic Demography at Lund University, individuals surviving to at least 1947 and their descendants have been linked to national registers from Statistics Sweden (*Statistiska centralbyrån*), the National Board of Health and Welfare (*Socialstyrelsen*) and the Military Archives (*Krigsarkivet, Pliktverket*), providing data also from 1968 up to present times.

SEDD is unique in several respects. It covers a longer period than most comparable databases and has a wealth of information at varying levels of aggregation. At the individual level, SEDD combines various demographic and socioeconomic records, including causes of death, place of birth and geographic data on the place of residence within a parish. The entire population present in the area is fully covered after 1813. For the city of Landskrona, the data cover the period 1905–1967, but the aim of an ongoing project is to complete the data entry back to 1880. At the family level, the data contain a combination of demographic records and information on occupation, landholding and income.

The multigenerational perspective of SEDD and the detailed information on a large number of variables are of great value to improve our knowledge of societal change in a broad sense. That SEDD covers the full period during which Sweden underwent enormous structural transformations, going from a preindustrial rural society to a modern welfare state, further underlining its scientific potential. SEDD is part of the Swedish national research infrastructure SwedPop, which aims to publish harmonized data from five different historical population databases using a common data structure (www.swedpop.se).

In this article, we first describe the source material and the linkage methods and then outline the data structure and variables in the database. The description covers the historical data up to 1968, and not the data from the national registers thereafter. Moreover, the data before 1813 will be covered in less detail as they are not included in the main database at present, and not structured in exactly the same way. These data are based on family reconstitutions, linked to poll-tax registers, rather than on population registers covering the entire resident population. A summary of the source material and the structure of the SEDD database is shown in Figure 1.

Figure 1 Structure of SEDD

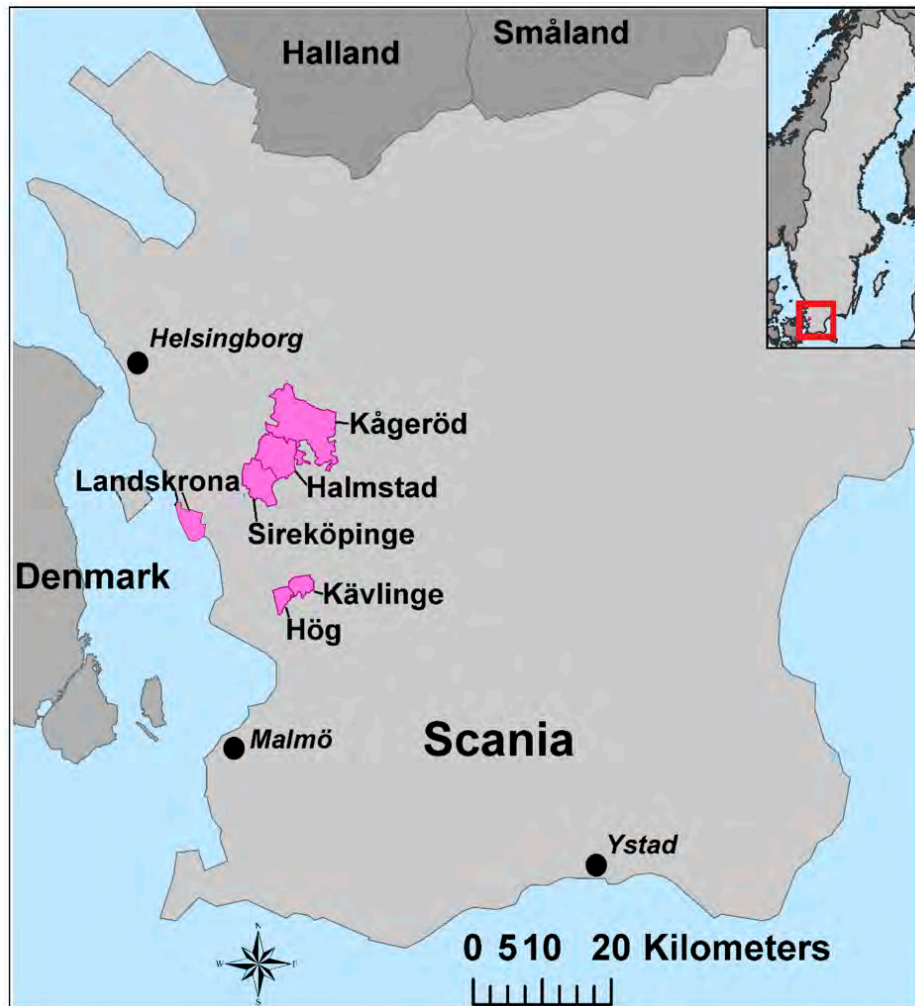


Source: Partly based on Figure 1 in Quaranta (2015).

2 THE AREA

SEDD consists of data from five rural and semi-urban parishes (Halmstad, Hög, Kågeröd, Kävlinge and Sireköpinge) and a port town, Landskrona, in Southern Sweden (Bengtsson, Dribe, Quaranta, & Svensson, 2020). The five parishes had a combined population of 4,300 in 1850, 5,900 in 1900 and, together with Landskrona, about 32,000 in 1950 (see figure 2 and table 1). For more detailed descriptions of the area and its socioeconomic and demographic development, see Bengtsson (2004), Dribe (2000), and Dribe and Svensson (2019).

Figure 2 Map of SEDD areas



Source: Map by Finn Hedefalk, Lund University.

Table 1 Population totals in the parishes in SEDD

	Hög	Kävlinge	Halmstad	Sireköpinge	Kågeröd	Landskrona
1750	208	193	312	360	1,274	
1805	292	344	529	613	1,412	
1850	550	631	787	817	1,544	4,139
1900	453	1,755	780	1,287	1,650	14,076
1950	365	3,208	348	881	1,796	25,089
1990	256	6,122	169	729	1,964	26,472

Sources: 1750–1850: *Tabellverket*, Umeå University (<http://rystad.ddb.umu.se:8080/Tabellverket/Tabverk/>); 1900: *BiSOS A*, part 2, tab. 1; 1950: *Statistics Sweden 1977 (Sveriges officiella statistik), Folkmängd 31 december 1950–1975*; 1990: *Statistics Sweden 1991 (Sveriges officiella statistik), Folkmängd 31 december 1990, del 1–2*.

The parishes were chosen so that a geographically compact area could be obtained. This enable a study of the economic-demographic interactions without introducing biases stemming from, for example, regional differences. On the other hand, it will be difficult, without further qualifications, to generalize the results to cover Sweden as a whole.

Halmstad, Sireköpinge and Kågeröd are neighboring parishes, and were completely dominated by manorial (noble) land. They are located in a rather hilly part of the area, on the border between the agricultural plains and the more forested area of the northwest. Sireköpinge was plain land, Halmstad brushwood with some plain lands in the south and wooded areas in the north, and Kågeröd was more of a forest region. Hög and Kävlinge, on the other hand, are located on the plains, and here arable land dominated entirely, with virtually no wooded parts at all. In these two parishes landownership was dominated by freeholds and crown land. Thus, the topographical conditions as well as landownership differed considerably between the parishes. Towards the end of the 19th century, Kävlinge expanded considerably and was transformed into a small town as a result of the construction of the railroad and the establishment of a number of small industries.

Landskrona was founded in 1413 as a mercantile port town with a deep, natural harbor (see [Dribe & Svensson, 2019](#)). Later in the 16th century it also became an important fortified military town, but soon lost most of that role after Sweden gained control of the province of Scania. From the mid-19th century, factories and financial institutions were established. Its development was similar to other industrial cities, exemplified by the emergence of newspapers, schools, a hospital, institutional poor relief and old age care, a municipal board and governance, and a connection to the railway lines. The port was used for shipping grain from its hinterland, supporting the region's role as the country's breadbasket, and the last quarter of the 19th century saw the founding of mechanical factories and a shipyard, the latter of which would come to play an important role for the economy and also the identity of the town for nearly a century. In 1900, the population was about 14,000 and ranked as the tenth largest Swedish industrial city. The economic and demographic expansion continued until the industrial crisis of the 1970s, which had a large impact on the city.

3 CONTENT OF THE DATABASE

3.1 SOURCE MATERIAL

The core source material on which SEDD is based varies somewhat over time:

1680–1812 (only the five parishes):

- Family reconstitutions based on vital events (births, deaths, marriages)¹
- Poll-tax registers (from 1766 annual coverage, before that year only partial)

1813–1967 (for Landskrona 1905–1967):

- Catechetical examination registers/parish books — population registers²
- Vital events registers (births, deaths, marriages)
- In- and out-migration registers
- Poll-tax registers (until 1945)

1862–1967 (for Landskrona 1905–1967):

- Income and taxation registers

In the following sections, we briefly discuss these different source materials. For parts of the database, information on heights, health and cognitive ability have been linked from muster rolls ([Öberg, 2014](#))

1 In the earliest registers data on births come from baptisms and data on deaths come from burials. Data on reported births and deaths are available from 1685 in Hög and Kävlinge, 1710 in Halmstad and Sireköpinge, and from 1730 in Kågeröd.

2 The starting dates for the catechetical examination registers varies somewhat between the parishes: Kågeröd: 1813; Halmstad and Sireköpinge: 1821; and Hög and Kävlinge: 1829.

and detailed information of mother's and child's health from midwives' reports and contemporary birth registers (Lazuka, 2017). These sources will not be described in detail here but specific documentation is available for users of SEDD.

3.2 FAMILY RECONSTITUTIONS³

Family reconstitutions were carried out for the five parishes for the period 1646–1895. The method of family reconstitution has been widely used in historical demographic research since the 1950s, especially to reveal demographic patterns during periods for which no censuses or population registers are available. Furthermore, family reconstitutions are also a highly valuable complement to aggregate demographic data, when trying to reveal demographic patterns at the micro level (e.g. Lundh & Bengtsson, 1989). By linking information on births (sometimes baptisms), deaths (sometimes burials) and marriages, whole families can be reconstructed (Alter, 2020). Since the method is very laborious and time consuming, it is difficult to cover large areas. Hence, family reconstitution studies often cover only one or a few parishes.

Several problems and limitations of the method have been identified. Firstly, the church records suffer from a certain degree of under-recording of vital events. However, the Swedish marriage records seem to compare fairly well with those of some other European countries, which is mainly attributable to its religious homogeneity, which meant that most people married in the state church and were registered in the church records (Lundh & Bengtsson, 1989). However, regarding infant deaths there seems to be a problem of under-recording in the Swedish registers, so that some children that died shortly after birth were neither registered as born nor as dead (Bengtsson, 1999; Bengtsson & Lundh, 1994). Naturally, this would lead to underestimation of both marital fertility and infant mortality.

Secondly, there is a problem of tracing families that move between parishes. Such families will spread their demographic events in different parishes, which makes a correct estimation of their time of exposure difficult. Often, the solution to this problem has been to limit the analysis, of for example fertility, to the families that can be completely reconstructed, i.e. those that married in the parish and then stayed there for the whole of their reproductive period (Alter, 2020). Due to the rather high mobility in preindustrial society, this implies that only a minority of the families can be used; sometimes as few as 10–15% of the total population. Furthermore, there has been concerns that the families that did not move for such long periods were not representative of the population as a whole (e.g. Åkerman, 1977; Hollingsworth, 1969; Thestrup, 1972). In practice, however, the effect of this type of selectivity bias seems to be rather limited (e.g. Levine, 1976; Rogers, 1988). It has also been argued that even if 'stayers' really were representative of the population as a whole, estimates based on family reconstitutions could be biased, since the probability of moving before marriage was highest for those that married late. Hence, late marriages are likely to be under-recorded due to migration (Ruggles, 1992). Similarly, the probability of moving before death would be higher for those that live longer. Therefore, they will be under-represented in the population, which in turn would lead to underestimation of life expectancy. However, regarding age at marriage in England, Wrigley (1994) argued that the problem in practice was less serious than Ruggles claimed (see also Ruggles, 1999). In SEDD, where family reconstitutions are supplemented with data on migration, place of residence etc., this problem is much less of a concern.

Thirdly, there is also a problem of identifying the individuals and correctly linking them to the right family. In SEDD, a computer-based linking procedure, using standardized names, was used for the family reconstitutions (Bengtsson & Lundh, 1993). Overall, the performance of the program was quite satisfactory: 96% of the births, 72% of the marriages and 55% of the deaths were correctly linked by the program (Bengtsson & Lundh, 1991). Since then, manual corrections raised these figures considerably (99% of births, 100% of marriages and 90% of the deaths were linked to families for the period before 1800).

3.3 POLL-TAX REGISTERS

Poll-tax registers (*mantalslängder*) are available from the late 17th century until 1945 and have been used to obtain information on where the families lived, and whether they had access to land or not. They were yearly registers, used for collecting taxes and containing information on the size of the

³ The sections on family reconstitutions, poll-tax registers and population registers are based on Dribe (2000, chapter 2).

landholding, the type of ownership (i.e. noble, crown, church or freehold) and information on the number of servants and lodgers. In addition to the poll-tax registers, land registers (*jordböcker*) have also been utilized to clarify the ownership of land, when poll-tax registers were missing. Information from these two registers was linked to the reconstituted families. Thereby, information was obtained not only on the demographic events, but also on the economic realities of these families. After 1766 the registers are available on an annual basis and linkage between the poll-tax registers and the family reconstitution is almost entirely complete (i.e. only odd holdings that have not been linked to a family).

In the poll-tax registers, the size of the landholding is expressed in *mantal*. This was an old tax unit, originally meaning 'the number of men'. At the beginning, during the 16th century, every landholding was supposed to have one *mantal*, i.e. be large enough to support one peasant and his family as well as producing a surplus to be paid as tax to the crown (Hechscher, 1949). Thus, at this time a *mantal* only meant that the peasant had land and was supposed to pay tax to the crown. However, due to repeated subdivisions of landholdings, farmsteads typically got smaller and smaller fractions of a *mantal* assigned to them. Furthermore, reclamation of new land, as well as changed methods of cultivation lead to increased land productivity, which makes a comparison over time of the size of different farms almost impossible. Nevertheless, the *mantal* can be used at least as a rough measure of the size of a farm relative to other farms in the village at the same point in time. Thus, by comparing the different *mantal* peasants had, the relative sizes of the landholdings could be determined.

The *mantal* assigned to a farm was based on the quantity of seed sown and hay harvested, which in turn reflected the productive potential of the arable land and of the meadow, even though other concerns were taken as well, such as the productive strength of underlying crofts or cottages (Sommarin, 1939). In western Scania in 1820, one *mantal* corresponded approximately to 180–230 acres (150–190 *tunnland*). At the same time, the minimum amount of land required to be a self-sufficient farmer (*besuttenhetsgräns*) was 1/16 of a *mantal*, which corresponded roughly to 15 acres (Sommarin, 1939).

Besides assessing the economic status of families, the poll-tax registers play a crucial role in determining the time of exposure for the families. By linking the reconstituted families to the place where they lived, we get the time they spent in the parish, and thereby we also get a rather precise estimate of time that these families are under observation. This information is of great use particularly for pre-1813 data, before population registers were available. It is important to note, however, that the information refers to the family, and not to individuals.

3.4 POPULATION REGISTERS (CATECHETICAL EXAMINATION REGISTERS AND PARISH BOOKS)

During the 18th century, a regulation concerning examinations of the biblical knowledge of the parishioners was introduced. To begin with, the examinations were held in church, by the people living in the parish coming to church (*kyrkoförhör*). Later, the examinations were instead conducted in people's homes (*husförhör*). The clergymen needed a list of the inhabitants in order to conduct the examinations, which was the origin of the catechetical examination registers. With the establishment of the Tabular Commission (*Tabellverket*) in 1749, the catechetical examination registers became the basis for the parish level records (Lext, 1984). This made it compulsory to record information on births, deaths and migration in the registers. In addition, the registers sometimes included information on reading and writing ability, smallpox vaccination, and various comments made by the clergy regarding the inhabitants. In the poll-tax instruction of 1812, the catechetical examination registers also got a very important role in controlling the poll-tax registers (Lext, 1967, 1984).

Every head of household was obliged to report the members of his household and the people living at his holding at the time of examination. Over the 19th century, the connection between the actual catechetical examinations and the catechetical examination registers diminished, and the clerical role of the registers disappeared to a large extent. Instead the demographic and administrative role of the registers became increasingly important (Lext, 1984), but the registers were still the responsibility of the parish. From 1895, the registers were renamed parish books (*församlingsböcker*) but contained the same basic information as before. This system of population registration was in place until 1991, when it was transferred to the tax authorities and became centralized.

Overall, there is a high degree of correspondence between the events in the catechetical examination registers and the records of vital events. However, children that died in infancy sometimes appear to

have been poorly registered in the catechetical examination registers, because not all children were entered immediately upon birth (see also [Winberg, 1975](#)).

In the population registers, the families and households were recorded according to where they lived. People belonging to the same household were listed together. Usually, it is relatively easy to identify to which household the head of household's family and his servants belonged based on where they are entered in the registers. However, it is sometimes more difficult to identify the right household for lodgers and retired people.⁴ If they are registered together with a certain family at a farm they have been included in that household, but in those cases where it has been difficult to determine to which household they belonged, they have been registered in a household of their own. As has been pointed out by other researchers, it is impossible to use the catechetical examination registers as a basis for a theoretical definition of a household ([Gaunt, 1977](#)). The definition used in SEDD focuses on the household as a production and consumption unit, and includes servants, lodgers and retired people in addition to the nuclear family of the head of household. This means that all the people living at the farm, contributing to production and/or dependent on the household for consumption, are included in the household, while temporary salaried workers that did not live in the household are not included. From 1947 and onwards, the population registers include unique personal identifiers given to all Swedish residents (*personnummer*).

3.5 INCOME AND TAXATION REGISTERS

For the period 1862–1967, SEDD contains information about income and taxes paid, derived from the income and taxation registers (*inkomst- och taxeringslängderna*). For the five parishes the information is available annually for the whole period 1862–1867, while for Landskrona it is available annually from 1947–1967, and about every five years from 1905–1946. Work is currently ongoing to complete the registration of income registers for all years between 1904 and 1946. While incomes initially were assessed by the local tax authorities, from 1902 onwards it was based on tax returns made by the tax payers themselves (*självdeklarationer*). The exact information included in the registers varies substantially over time, and changing income thresholds in the tax law implies that an increasing share of the population is included in the registers. [Helgertz, Dribe, and Bengtsson \(2020\)](#) provide a detailed description of the source material and the income variables included in SEDD.

Until 1971, Sweden practiced joint taxation of married couples, but after 1947 income of married women were nonetheless reported separately in the registers. Before 1947, married women are not included in the tax registers, but their incomes are reported together with that of their husbands.

3.6 LINKS TO THE SWEDISH DEATH INDEX AND THE CENSUS OF 1950

SEDD has been linked to both the Swedish Death Index (SDI) and the Census of 1950. The SDI has been developed by the Federation of Swedish Genealogical Societies ([2019](#)) and includes most deaths in Sweden between 1860 and 2017. After 1947, the index also includes the unique personal identification numbers which allow for a direct link to SEDD. Before 1947, the sources were linked using automated matching based on names and date and place of birth. The links are valuable in providing information about the time and place of death for out-migrants, as well as personal identification numbers for people moving out before 1947 (when we have the personal identifiers in SEDD), but dying after 1968 so that they could be traced in the national registers before they died.

In a similar way, the Census of 1950 (digitized and transcribed by Arkiv Digital, www.arkivdigital.se) was linked to SEDD using either personal identification numbers or through automatic matching based on names and date and place of birth. The census provides snapshot information from 1950 on occupation, family context and place of residence and, more importantly, gives personal identification numbers for people leaving the area before 1947, allowing more SEDD individuals to be linked to the national registers.

4 Since leasing agreements are not fully captured in the poll-tax registers, it is difficult to know which families lived at the same farm as each other and thus shared environment. For example, when several smaller units are listed under the main farm it can be difficult (but possible using historical maps) to link the correct tenant to the unit they rent, but even more difficult to link other families in the poll-tax registers that resided in either the rented unit or in the 'main' property unit, but in different households than the tenant/owner.

3.7 FATHER'S OCCUPATION AT BIRTH OF THE CHILD

All adults present in SEDD have been traced back to their parish of birth to retrieve information on the occupation of the father. The information has been taken from the birth register, or in some cases the parish book. This means that the database contains information on father's occupation not only for children born in the parishes, but for all individuals who were present in SEDD as adults and for whom that information could be found in the parish of birth.

4 DATA CODING AND LINKAGE BETWEEN SOURCES

4.1 CODING OF OCCUPATIONS

All occupations have been coded into HISCO, the *Historical International Standard Classification of Occupations* (van Leeuwen, Maas, & Miles, 2002). HISCO is a historical classification based on the International Labour Organisation's classification ISCO68 (ILO, 1969) and attempts to increase the comparability of historical occupational titles. In HISCO occupations are categorized according to tasks that need to be fulfilled in that occupation. HISCO divides occupations into eight major groups (e.g. major group 5 'Service Workers'), each of which is divided in two to ten minor groups (e.g. minor group 5.3. 'Cooks, Waiters, Bartenders and Related Workers'). These 83 minor groups are again subdivided into 284 unit groups (e.g. 5.31 'Cooks'). Finally, these unit groups consist of 1,881 occupational categories, the lowest level of detail (e.g. 5.31.50 'Ship's Cook') (see Dribe, Helgertz, & Van de Putte, 2015). Occupations with comparable tasks are grouped into one of these categories. Initially the coding was done manually, but in the present version of SEDD the occupational coding done in the SwedPop project has been applied to the SEDD data. This coding is a harmonized version of HISCO coding done at different historical population databases in Sweden. Based on HISCO codes, social class schemes, such as HISCLASS (Historical International Social Class Scheme, see van Leeuwen & Maas, 2011), and rank schemes, such as HISCAM (Lambert, Zijdemann, van Leeuwen, Maas, & Prandy, 2014), can be easily applied using existing transcode tables.

4.2 CODING OF CAUSES OF DEATH

SEDD includes information about cause of death. The unique text strings indicating the causes listed in the death registers have been coded into ICD10 (*International Statistical Classification of Diseases and Related Health Problems*) within the SwedPop project. As with the HISCO codes, this is a harmonized coding of causes of death applied in all the large historical population databases in Sweden (Hiltunen & Edvinsson, 2018).

4.3 DATA LINKAGE 1813–1967

The data from the different sources have been linked together manually, in most cases when a new source was added to the database. Information about name, date of birth, place of birth, and family context have been important in identification and linking of records. The vital events registers have also been linked to the population registers to check for under-recording of events, and to add missing data when necessary.

As already mentioned, linking of SEDD to the census of 1950 and the SDI, as well as the original family reconstitution of the five parishes, was done wholly or partly using automatic linkage methods.

Table 2 shows the number of events and unique individuals in the database 1813–1967. In total, it includes about 175,000 unique individuals, of which 74,000 have personal identification numbers that allow them to be linked to the national contemporary registers. The database contain about 43,000 births, 27,000 deaths and almost 300,000 migrations, out of which about 20,000 moves are between the parishes in the database.

Table 2 *Number of events, individuals, families and households in SEDD, 1813–1967*

	N
Births	43,129
Deaths	26,676
Marriages	26,171
In-migrations	148,831
Out-migrations	145,755
Number of individuals	175,149
Individuals with national id number	74,213
Number of family IDs	116,888
Number of household IDs	79,744

Note: In/out migrations are moves over parish borders, also including migrations between two parishes in the sample.

4.4 GEOCODING⁵

Through geocoding, the place of residence is known for parts of the population in the five parishes and Landskrona at various geographic levels and periods.

For the five parishes, the approximately 53,000 residents⁶ for the period 1813–1914 have been linked to the property units at which they lived (see Hedefalk, Harrie, & Svensson, 2015). To enable such geocoding, an object-lifeline representation of the digitized property units was created (using historical maps and poll-tax registers), which contains information about when the units were created, changed, and ceased to exist. Such a representation was necessary to accurately trace the residential histories in time, as the property unit borders were often subdivided or partitioned into smaller units in line with the rapid population growth during the period. Thus, each move within the parishes can be traced. In addition to the geocoding, historical wetlands (in object-lifelines), roads, buildings and water in the five parishes were digitized for the same period (see Hedefalk et al. (2017) for details on the geocoding match rate and datasets).

For Landskrona, the full population has been geocoded at: (1) block-level for the period 1904–1938; and (2) address-level for the period 1939–1967. Thus, continuous information about the individual place of residence at a very detailed geographic level is available. For the period 1939–1967, individuals are linked to the exact addresses they lived at, and each move is traced. For the period 1904–1938, the geocoding has been performed on the slightly coarser block-level (Swedish: *kvarter*). In addition, an object-lifeline representation of the buildings and streets in Landskrona was created for the period 1904–1967, allowing to link individuals to the buildings they resided in (more accurately for the period 1939–1967; see Hedefalk and Dribe (2020) for information on geocoding match rate).

5 We thank Finn Hedefalk for contributing to this section.

6 For the geocoding of the five parishes, we primarily used information from the poll-tax registers, which contains the addresses (property units) of each person who had to pay taxes. The poll-tax registers are on a household level, which indicates that only the head of the household is noted with a full name. However, each individual in the SEDD is linked to the families and households to which they belong, and most of these families and households have a correspondence in the poll-tax registers.

5 DATA STRUCTURE AND VARIABLES

5.1 FROM SOURCE MATERIAL TO A DATASET FOR ANALYSIS

The linked data from the different source registers is stored in a relational database in Microsoft SQL server: the SEDD source database (see Figure 1). This database is composed of different tables containing information from each source, and tables with standardized variables (occupations, locations and causes of death). The SEDD source database is continuously updated as new data is digitized or any information is changed.

The data for 1813–1967 was structured in the model of the Intermediate Data Structure (IDS).⁷ The IDS was introduced with the aim of creating a common format for the standardization and dissemination of data from different databases, regardless of their original form (Alter & Mandemakers, 2014). It also provides standardized solutions for storing constructed variables, making data extractions and preparing datasets for analysis, and allows to develop and share common software (Quaranta, 2015). The IDS is not meant to replace source databases, but rather to serve as a middle layer between the source database and rectangular datasets for analysis.

IDS databases are composed of two types of entities, persons and contexts, and the relation among persons, contexts and between persons and contexts (Alter & Mandemakers, 2014). Contexts represent physical or social spaces that group individuals together. Data is stored in the IDS tables following the entity-attribute-value model, which contains one 'attribute' (i.e. variable) per record. Each record includes a *Type* or *Relation*, indicating the kind of information it contains, which can be the date of an event or *Value* of an attribute, or the type of relationship that exists between two persons or between a person and a context (Alter, Newton, & Oeppen, 2020). Each record also includes an identifier of the person (*Id_I*) or context (*Id_C*). Attributes are dated using a *Time Stamp*.

To conduct longitudinal statistical analysis using data stored in IDS, data extractions must be transformed into rectangular data arrays, also called 'episodes tables'. Episodes are spells of time during which the values of attributes remain constant and at the end of which the event of interest of the study can take place. The start and end dates of the rows of an episodes table correspond to the dates when any of the attributes or events included in the data extraction change value. The transformation of IDS data into episodes tables can be done using the two-step procedure proposed in Quaranta (2015). In the first step, a Chronicle and a Variable setup file are produced. The Chronicle file contains all individual and contextual level attributes and all events selected by the researcher for analysis. Each *Type* of attribute included in the Chronicle file becomes a column of the episodes table. In the second step, episodes tables can be automatically made from these two files, using the STATA program 'Episodes file creator' (Quaranta, 2016). The Variable Setup file stores information relating to each attribute included in the Chronicle file, and this information is used by the 'Episodes file creator' to identify how each attribute should be treated when creating the episodes table.

5.2 THE SEDD IDS TABLES

The IDS consists of five main tables: INDIVIDUAL storing information relating to individuals; INDIV_INDIV defining relationships between individuals; CONTEXT defining contexts and storing information about them; CONTEXT_CONTEXT defining contexts that are nested in other higher level contexts; and INDIV_CONTEXT identifying spells of times during which individuals are present in a specific context (Alter & Mandemakers, 2014). A METADATA table is also used, which defines all attributes included in the tables and their values.

The SEDD INDIVIDUAL table includes many attributes. Dates of birth and marriage are available, and for such records the *Date_type* field of the table specifies whether the actual birth and marriage events were observed or whether their dates were declared in other sources. Birth and marriage locations and information on death date, location and cause of death (ICD10 codes) are also stored in the table.

7 Separate IDS tables were made for the family reconstitution data, which include source information from the vital event registers and the poll-tax registers. Family reconstitution rules should be adopted when using this data for research.

Additional variables include sex, arrival, departure and start and end observations.⁸ The INDIVIDUAL table also stores occupations of individuals, indicating in the *Source* field of the table which register the occupation comes from (parish, or income and taxation registers), and the occupation of the father at the time of birth of the individual. In both cases, standardized strings and HISCO coding are available. The types of relations included in the SEDD INDIV_INDIV table are husband, wife, child, father, mother, adoptive child/father/mother, foster child/father/mother and step child/father/mother.⁹

In the SEDD, four different levels are included in the CONTEXT table: family, household, parish and poll-tax.¹⁰ The family includes all related persons within one household. It corresponds to the nuclear family, also referred to as conjugal family unit (CFU — see e.g. Hammel & Laslett, 1974). CFUs may be formed in any of the following ways: by a couple without offspring; by a couple with unmarried offspring and/or unmarried adopted/foster children; by a lone parent with at least one never-married child; or by a single adult. Families never include more than two generations. If more than two generations live in the same household, the CFU is formed from the youngest generation upwards.¹¹ Servants and lodgers constitute their own families. Individuals are linked to family contexts in the INDIV_CONTEXT table, specifying the type of relation that they have to the family (man, woman, child). Individuals can only belong to one family at any specific point in time, but they can move across families (e.g. family of birth and family of marriage).

As discussed above, households in the SEDD include the nuclear family of the head of household as well as servants, lodgers and retired people. Several families may live in the same household. For example, a household will host the main family (which is the head of household, his wife and children), plus a family of servants, consisting of a married couple and their children. Individuals are linked to a household context in the INDIV_CONTEXT table, specifying the type of relation that they have to the household (householder, relative, boarder, servant or unknown). Individuals can only belong to one household at any specific point in time, but they can move across households. In SEDD families are not nested within households since it is possible for individuals who belong to the same family to live in different households at the same point in time.¹² Households are linked to parishes through the CONTEXT_CONTEXT table.

A specific context layer was created for poll-tax, to allow to more easily link information from these registers to all members of the family. The variables from poll-tax registers that are stored in the CONTEXT table are holding and land type, size fraction (*mantal*) and occupation. Poll-tax registers are linked to families through the CONTEX_CONTEXT table. Since families could own more than one property, it is possible for them to have more than one link to a poll-tax register during the same year.

5.3 THE SEDD EXTENDED IDS TABLES

In addition to storing source data in the SEDD IDS tables, a wide range of individual- and context-level variables were constructed using source information, like household size, total number of children in the family, occupation of the family head, etc. They were stored in the extended IDS tables (EIDS — Quaranta, 2015) INDIVIDUAL_EXT and CONTEXT_EXT. Most types of analyses using SEDD can therefore be conducted without the need of further elaboration of the data. This not only facilitates the work of researchers, but it also makes different research outputs more consistent, since variables are defined identically.

- 8 All dates of migration to and from SEDD areas and within SEDD areas are recorded. When it was possible, individuals who moved between different SEDD parishes were linked and given unique identifiers. Such links could be made for all individuals after 1947, when personal identification numbers became available. In preceding years, links were established if there was a full match on name and date and place of birth; however, this procedure is unlikely to have captured all cases.
- 9 The relations stored in the INDIV_INDIV table correspond to those found in the source material. Other types of kinship relations can be identified using the information from this table. For example, if individual 1 is the mother of individual 2, the grandchildren of 1 can be identified by selecting from the table all children linked to 2.
- 10 In the family reconstitution IDS dataset, the CONTEXT table includes the levels family and poll-tax, but not household, given that population registers are not available.
- 11 Other extended family members, such as siblings of the head of family also constitute their own separate family.
- 12 An example can be a married woman spending a period as lodger in a different household and later returning to the household where the rest of her family lives.

The CONTEXT_EXT table stores constructed context-level variables. Two of such variables are family head ID and household size. The occupations of the family head are also included in the tables, which correspond to the occupations, from the parish registers and from the income and taxation registers, of the individual defined as the family head. The source field of variables is not retained when constructing an episodes table. However, because of the importance in research to distinguish occupations from different sources, all created extended occupational variables incorporate the source into the variable name. Amongst such created variables are the occupations from the poll-tax registers, and since families can have more than one poll-tax declaration each year, up to three poll-tax occupations may be available. Three variables indicating the date when occupations were last declared are also stored in the table; one for the occupations of the family head from the parish registers, one for the occupations of the family head from the income and taxation registers and one for the occupations from the poll-tax registers. When constructing the episodes tables, values of occupations are copied down until the next time an occupation is declared. Using the declaration date variables, researchers can later decide for how long to consider occupations valid.¹³

The INDIVIDUAL_EXT table currently contains more than 100 variables, which can be divided into different groups. One group of variables measures characteristics of the family, some of which only relate to children (mother presence, father presence, older sisters, older brothers, younger sisters, younger brothers), others only to the man and the woman (number of children), and others to all members of the family (proportion of previously born children who are dead, number of surviving daughters, number of surviving sons). There is also one general individual level variable, civil status, which takes into consideration whether the individual has a partner, and whether the partner is present in the same household.

The variables stored in the CONTEXT_EXT table (household size, size fraction, holding type, land type, poll-tax occupations and occupations of the family head) could also be assigned to individuals for the specific periods of time in which they were present in the relative context. Such information is stored in the INDIVIDUAL_EXT table as well. This makes this information redundant in the database, but it has the advantage that it can more easily be used in research. Occupations of the individual, from the parish registers and from the income and taxation registers, are also stored in INDIVIDUAL_EXT, incorporating also in this case the source information into the name of the variable. Since each occupational title may contain up to four occupations, for each occupational variables there are twelve variables related to HISCO coding (code, relation and status). Moreover, the table stores variables with the dates of declaration of each type of occupational variable (see above).

Finally, the INDIVIDUAL_EXT table also stores a series of variables that are specific to mortality or fertility research, and a variable measuring whether the individual was present in the study area, which is essential for most kinds of longitudinal research. For mortality research, it includes an indicator variable of whether the individual died at the end of the spell. For fertility research, the table includes several variables, which are assigned only to females: indicator of child birth, date of previous child birth, date of previous marriage, number of previous births, status of the previously born child.

13 For example, the occupation of the family head, from the parish registers, could have the value farmer, declared on March 13, 1856, when the family first enters the study area, the value shoemaker, declared on August 5, 1862, and no other occupational declarations until the family leaves the study area on January 21, 1865. When creating the episodes table the value farmer is assigned to the spell between March 13, 1856 and August 5, 1862, and the value shoemaker is assigned to the spell between August 5, 1862 and January 21, 1865. Since there are many variables in the database, which change values in different dates, each spell between declarations of occupations is generally split into more rows, making it difficult for researchers to know when an occupation was actually declared. This is why the variable indicating the date of declaration of occupations becomes important. Using the date variable a researcher could for example decide to hold occupations valid for up to five years and therefore replace the occupation value to missing between March 13, 1861 and August 5, 1862.

6 USING SEDD IN RESEARCH

In addition to the IDS and EIDS tables, a Chronicle file and a Variable setup file are provided to researchers.¹⁴ The SEDD Chronicle file contains all source and constructed variables and events. As mentioned earlier, the Chronicle and Variable setup files can be automatically transformed into a rectangular episodes table that is ready for statistical analysis using a STATA program (Quaranta, 2016). The episodes table can also be exported to conduct analysis using other software. In addition to the variables provided, researchers can create other variables using information contained in the IDS tables or external data. Such variables should be included in the Chronicle and Variable setup files prior to creating the episodes table.

The transformation of the SEDD source database into IDS and the construction of variables for analysis is made using SQL queries in Microsoft Access and Microsoft SQL Server Management Studio. A new IDS database and new Chronicle and Variable setup files are produced every time there is a new SEDD release. New releases are made when sections of digitization are completed, or when substantial errors are fixed in the data.

Data from SEDD older than 100 years are publicly accessible without restrictions.¹⁵ More recent data pertaining to individuals who are still alive can be released for scientific research subject to legal restrictions on the use of personal data (e.g. GDPR and the Swedish law of ethical review of research).

7 SUMMARY AND FUTURE PERSPECTIVES OF SEDD

This article has described the SEDD database, its sources and structure. As was shown, SEDD contains detailed and rich longitudinal demographic and socioeconomic data, which expand over a very long time period and across several generations. The conversion of SEDD into IDS has meant that the structure of the database is clear and well documented, making SEDD easily available for researchers. Many variables have also been constructed and episodes tables can be automatically created, meaning that most analyses can be made without additional elaboration of the data.

Even though SEDD is very rich, there are several plans to continue its expansion. As mentioned earlier, the data for Landskrona will be extended back to 1880. More socioeconomic and health variables will be added to SEDD, expanding further the scope of the data. For example, hospital birth records are available for children born between 1935 and 1945 and more cohorts are currently being digitized. Like many other longitudinal databases, the main limitation of SEDD is its geographical coverage. The IDS has however allowed to conduct fully comparative studies across different European populations (e.g. Quaranta, 2018; Quaranta & Sommerseth, 2018), increasing the generalizability of results. SEDD is also part of the SwedPop project, which aims to harmonize and disseminate historical population data from different sources and areas of Sweden (www.swedpop.se). Swedpop will vastly expand the possibilities to conduct comparative research considering other areas in Sweden.

ACKNOWLEDGEMENTS

This paper has been produced within the research programme 'Landskrona Population Study', funded by the Swedish Foundation for Humanities and Social Sciences (RJ). SEDD is part of SwedPop which is a national infrastructure funded by The Swedish Research Council (VR) together with the partner institutions, Umeå University, Lund University, University of Gothenburg, the National Archives, and the Stockholm City Archives. We are grateful to Tommy Bengtsson and Finn Hedefalk for comments and suggestions.

14 In the IDS family reconstitution database, only basic IDS tables were made, i.e. no extended variables were constructed and the Chronicle and Variable Setup file are not provided.

15 For more information, see <https://www.ed.lu.se/databases/sedd/sedd-public-access>.

REFERENCES

- Åkerman, S. (1977). An evaluation of the family reconstitution technique. *Scandinavian Economic History Review*, 25, 160–170. doi: [10.1080/03585522.1977.10407879](https://doi.org/10.1080/03585522.1977.10407879)
- Alter, G. (2020). The evolution of models in historical demography. *Journal of Interdisciplinary History*, 50(3), 325–362. doi: [10.1162/jinh_a_01445](https://doi.org/10.1162/jinh_a_01445)
- Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal microdata, version 4. *Historical Life Course Studies*, 1(1), 1–29. Retrieved from <http://hdl.handle.net/10622/23526343-2014-0001?locatt=view:master>
- Alter, G., Newton, G., & Oeppen, J. (2020). Re-introducing the Cambridge Group Family Reconstitutions. *Historical Life Course Studies*, 9, 24–48. Retrieved from <https://hdl.handle.net/10622/23526343-2020-0005>
- Bengtsson, T. (1999). The vulnerable child. Economic insecurity and child mortality in preindustrial Sweden: A case study of Västanafors, 1757–1850. *European Journal of Population*, 15, 117–151. doi: [10.1023/A:1006215701608](https://doi.org/10.1023/A:1006215701608)
- Bengtsson, T. (2004). Mortality and social class in four Scania parishes. In: Bengtsson, T., Campbell, C., & Lee, J. Z. (Eds.). *Life under pressure: Mortality and living standards in Europe and Asia 1700–1900* (pp. 37–41). Cambridge, MA: MIT Press.
- Bengtsson, T., Dribe, M., Quaranta, L., & Svensson, P. (2020). *The Scania Economic Demographic Database. Version 7.1 [machine-readable database]*. Lund: Lund University, Centre for Economic Demography.
- Bengtsson, T., & Lundh, C. (1991). *Evaluation of a Swedish computer program for automatic family reconstitution*. (Lund Papers in Economic History, No. 8). Lund: Lund University, Department of Economic History.
- Bengtsson, T., & Lundh, C. (1993). *Name-standardisation and automatic family reconstitution*. (Lund Papers in Economic History, No. 29). Lund: Lund University, Department of Economic History.
- Bengtsson, T., & Lundh, C. (1994). Child and infant mortality in the Nordic countries. *Annales de démographie historique*, 24–43. doi: [10.3406/ADH.1994.1857](https://doi.org/10.3406/ADH.1994.1857)
- Dribe, M. (2000). *Leaving home in a peasant society. Economic fluctuations, household dynamics and youth migration in southern Sweden, 1829–1866*. Södertälje: Almqvist & Wiksell International.
- Dribe, M., Helgertz, J., & Van de Putte, B. (2015). Did social mobility increase during the industrialization process? A micro-level study of a transforming community in southern Sweden 1828–1968. *Research in Social Stratification and Mobility*, 41, 25–39. doi: [10.1016/j.rssm.2015.04.005](https://doi.org/10.1016/j.rssm.2015.04.005)
- Dribe, M., & Svensson, P. (2019). *Landskrona 1900–2000 — A comparative analysis of the economic and demographic development*. (Lund Papers in Economic Demography, No. 3). Lund: Lund University, Department of Economic History. Retrieved from https://www.ed.lu.se/media/ed/papers/working_papers/LPED_2019_3.pdf
- Gaunt, D. (1977). Pre-industrial economy and population structure: The elements of variance in early modern Sweden. *Scandinavian Journal of History*, 2(1–4), 83–210. doi: [10.1080/03468757708578918](https://doi.org/10.1080/03468757708578918)
- Hammel, E. A., & Laslett, P. (1974). Comparing household structure over time and between cultures. *Comparative studies in Society and History*, 16(1), 73–109. Available from <https://doi.org/10.1017/S0010417500007362>
- Heckscher, E. F. (1949). *Sveriges ekonomiska historia från Gustav Vasa, del 2*. Stockholm: Albert Bonniers förlag.
- Hedefalk, F., & Dribe, M. (2020). The social context of nearest neighbors shapes educational attainment regardless of class origin. *Proceedings of the National Academy of Sciences*, 117(26), 14918–14925. doi: [10.1073/pnas.1922532117](https://doi.org/10.1073/pnas.1922532117)
- Hedefalk, F., Harrie, L., & Svensson, P. (2015). Methods to create a longitudinal integrated demographic and geographic database on the micro-level: A case study of five Swedish rural parishes, 1813–1914. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 48(3), 153–173. doi: [10.1080/01615440.2015.1016645](https://doi.org/10.1080/01615440.2015.1016645)
- Hedefalk, F., Svensson, P., & Harrie, L. (2017). Spatiotemporal historical datasets at micro-level for geocoded individuals in five Swedish parishes, 1813–1914. *Scientific data*, 4(1), 1–13. doi: [10.1038/sdata.2017.46](https://doi.org/10.1038/sdata.2017.46)
- Helgertz, J., Dribe, M., & Bengtsson, T. (2020). *Income data in the Scania Economic Demographic Database (SEDD)*. (Lund Papers in Economic Demography, No. 6). Lund: Lund University, Department of Economic History. Retrieved from https://www.ed.lu.se/media/ed/papers/working_papers/LPED_2020_6.pdf

- Hiltunen, M. & Edvinsson, S. (2018). Classifying literate cause-of-death information originating from Swedish historical parish registers. Unpublished paper.
- Hollingsworth, T. H. (1969). *Historical Demography*. London: Hodder & Stoughton.
- International Labor Office (ILO). 1969. *International standard classification of occupations, revised edition 1968*. Geneva: International Labor Office.
- Lambert, P. S., Zijdeman, R. L., van Leeuwen, M. H. D., Maas, I., & Prandy, K. (2013). The construction of HISCAM: A stratification scale based on social interactions for historical comparative research. *Historical Methods*, 46(2), 77–89. doi: [10.1080/01615440.2012.715569](https://doi.org/10.1080/01615440.2012.715569)
- Lazuka, V. (2017). *Defeating disease. Lasting effects of public health and medical breakthroughs between 1880 and 1945 on health and income in Sweden* (PhD Thesis Economic History, Lund University). Retrieved from <https://lup.lub.lu.se/search/publication/1cba7526-042f-43fe-9005-6f6bc8cbf862>
- Levine, D. (1976). The reliability of parochial registration and the representativeness of family reconstitution. *Population Studies*, 30(1), 107–120. doi: [10.1080/00324728.1976.10412723](https://doi.org/10.1080/00324728.1976.10412723)
- Lext, G. (1967). *Mantalskrivningen i Sverige före 1860*. (Meddelanden från Ekonomisk-historiska institutionen, No. 13). Göteborg: Ekonomisk-historiska institutionen.
- Lext, G. (1984). *Studier i svensk kyrkobokföring 1600–1946*. (Meddelanden från Ekonomisk-historiska institutionen vid Göteborgs universitet, No. 54). Göteborg: Ekonomisk-historiska institutionen.
- Lundh, C., & Bengtsson, T. (1989). *Familjerekonstruktion på svenskt kyrkoboksmaterial. Problem och möjligheter*. (Meddelande från ekonomisk-historiska institutionen, Lunds universitet, No. 59). Lund: Ekonomisk-historiska Institutionen.
- Öberg, S. (2014). *Social bodies: Family and community level influences on height and weight, southern Sweden 1818–1968* (PhD Thesis Economic History, University of Gothenburg). Retrieved from https://www.researchgate.net/publication/281290743_Social_bodies_family_and_community_level_influences_on_height_and_weight_southern_Sweden_1818-1968
- Quaranta, L. (2015). Using the Intermediate Data Structure (IDS) to construct files for statistical analysis. *Historical Life Course Studies*, 2, 86–107. Retrieved from <http://hdl.handle.net/10622/23526343-2015-0007?locatt=view:master>
- Quaranta, L. (2016). STATA programs for using the Intermediate Data Structure (IDS) to construct files for statistical analysis. *Historical Life Course Studies*, 3, 1–19. Retrieved from <http://hdl.handle.net/10622/23526343-2016-0001?locatt=view:master>
- Quaranta, L. (2018). Intergenerational transfers in infant mortality in southern Sweden, 1740–1968. *Historical Life Course Studies*, 7, 88–105. Retrieved from <http://hdl.handle.net/10622/23526343-2018-0013?locatt=view:master>
- Quaranta, L., & Sommerseth, H. L. (2018). Introduction: Intergenerational transmissions of infant mortality using the Intermediate Data Structure (IDS). *Historical Life Course Studies*, 7, 1–10. Retrieved from <http://hdl.handle.net/10622/23526343-2018-0014?locatt=view:master>
- Rogers, J. (1988). *Family reconstitution: new information or misinformation*. (Reports from the Family History Group, 7). Uppsala: Department of History, Uppsala University.
- Ruggles, S. (1992). Migration, marriage and mortality: Correcting sources of bias in English family reconstitutions. *Population Studies*, 46(3), 507–522. doi: [10.1080/0032472031000146486](https://doi.org/10.1080/0032472031000146486)
- Ruggles, S. (1999). The limitations of English family reconstitution: English population history from family reconstitution 1580–1837. *Continuity and Change*, 14(1), 105–130. doi: [10.1017/S0268416099003288](https://doi.org/10.1017/S0268416099003288)
- Sommarin, E. (1939). *Det skånska jordbrukets ekonomiska utveckling 1801–1914, del 2–3*. Lund: Skrifter utgivna av de skånska hushållningssällskapen.
- The Federation of Swedish Genealogical Societies (2019). *Sveriges dödbok, version 7, 1860–2017* [Machine readable database].
- Thestrup, P. (1972). Methodological problems of a family reconstitution study in a Danish rural parish before 1800. *Scandinavian Economic History Review*, 20(1), 1–26. doi: [10.1080/03585522.1972.10407708](https://doi.org/10.1080/03585522.1972.10407708)
- van Leeuwen, M. H. D., & Maas, I. (2011). *HISCLASS. A Historical International Social Class Scheme*. Leuven: Leuven University Press.
- van Leeuwen, M. H. D., Maas, I., & Miles, A. (2002). *HISCO. Historical International Standard Classification of Occupations*. Leuven: Leuven University Press.
- Winberg, C. (1975). *Folkökning och proletarisering. Kring den sociala strukturomvandlingen på Sveriges landsbygd under den agrara revolutionen* (PhD Thesis History, University of Gothenburg).
- Wrigley, E. A. (1994). The effect of migration on the estimation of marriage age in family reconstitution studies. *Population Studies*, 48(1), 81–97. doi: [10.1080/0032472031000147486](https://doi.org/10.1080/0032472031000147486)

HISTORICAL LIFE COURSE STUDIES
VOLUME 9 (2020), published 14-12-2020

A Longitudinal Historical Population Database in Asia

The Taiwanese Historical Household Registers Database (1906–1945)

Chia-chi Lin	Tamkang University
Shu-juo Chen	National Museum of Natural Science
Ying-chang Chuang	Academia Sinica
Wen-shan Yang	Academia Sinica
James Wilkerson	National Tsing Hua University
Ying-hui Hsieh	Tzu Chi University
Ko-hua Yap	National Sun Yat-sen University
Yu-ling Huang	Academia Sinica

ABSTRACT

For the past 35 years, the Taiwan Historical Household Registers Database (THHRD) has been significant for historical demographic research on Asia. In recent years, researchers have continued adding new demographic information to the database. This allows for the expansion of research on the topic of historical households in the region. However, there are still many issues to address in the field of Asian historical demography. This paper provides a brief introduction on the uses of THHRD for future research.

Keywords: Taiwan Historical Household Registers Database, Historical demography, Life events, Taiwan

DOI article: <https://doi.org/10.51964/hlcs9300>

© 2020, Lin, Chen, Chuang, Yang, Wilkerson, Hsieh, Yap, Huang
This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

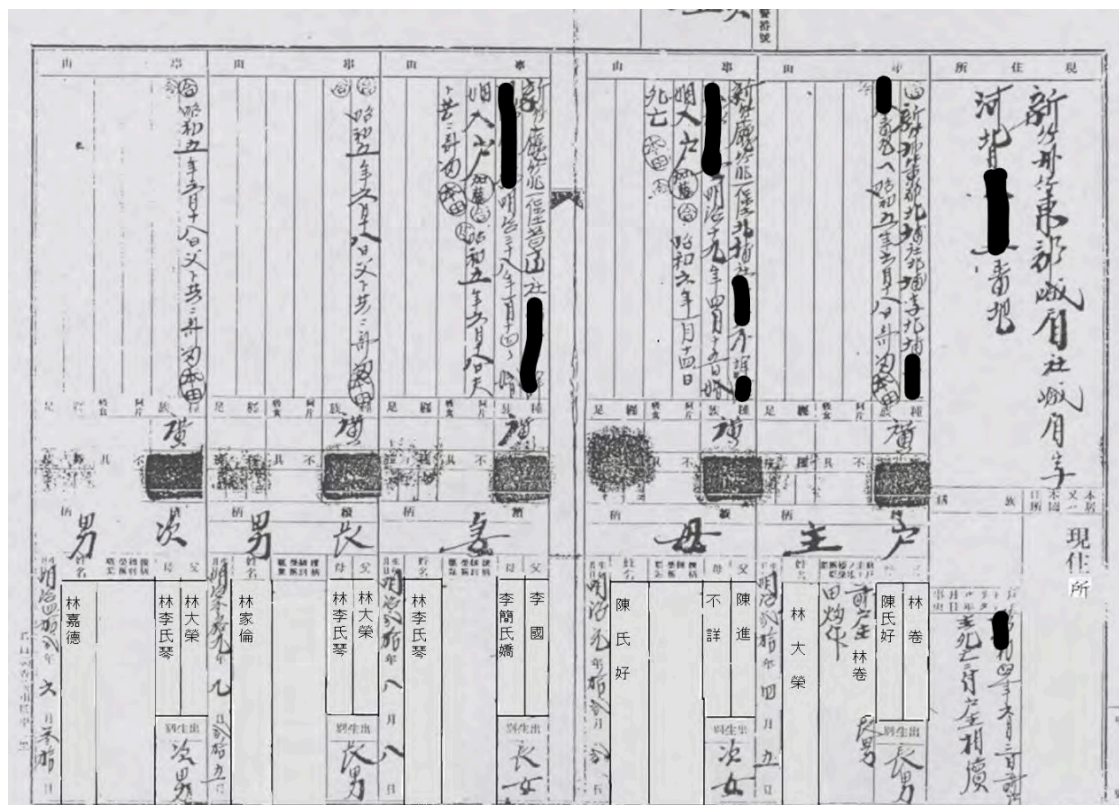
1 INTRODUCTION

Stanford anthropologist Arthur P. Wolf was the first scholar to recognize the academic value of the Taiwanese household registers. He utilized them to explore the marriage and adoption customs of the Han Chinese (Wolf, 1968; Wolf & Huang, 1980). Thanks to Professor Wolf we have some unpublished manuscripts about the Taiwan Historical Household Registers Database, 1906–1945 (THHRD), maintained by the Program for Historical Demography (PHD) at the Academia Sinica, Taipei, Taiwan. This paper is based on Professor Wolf’s 2009 manuscript, as well as recent developments and achievements.

In 1985, Professor Arthur P. Wolf began to cooperate with Ying-chang Chuang, Research Fellow of the Institute of Ethnology, Academia Sinica. Together they collected the Taiwanese household registers¹ compiled during the Japanese colonial era. They then digitized the information in these registers in order to construct a historical demographic database. In 2003, the Program of Historical Demography² was officially launched at the Research Center for Humanities and Social Science (RCHSS), Academia Sinica, to continue the project of digitizing the data of the Taiwanese household registers.

Taiwan was ruled by Japan from 1895 to 1945 after China was defeated in the Sino-Japanese War of 1894. In order to tightly monitor and control this newly seized island, the Japanese colonial government soon implemented a carefully designed household registration system (Figure 1). The language of the household registers is Japanese. It covered great details of the social life of individuals and families at that time and serves as a precious legacy left behind by the colonial authorities. 'One can recover a large part of the life history of every person alive in the period 1905–1945, and reconstruct the exact composition of each family he or she joined.' (Wolf, forthcoming). These detailed records allow for a variety of studies on demographic events from a longitudinal perspective and can make a contribution to many academic disciplines.

Figure 1 A household register of Colonial Taiwan



Source: *The Program for Historical Demography* (2020).

Note: Due to privacy protection regulations, the names in Figure 1 are censored.

- 1 These household registers are preserved in the household registry office in the located townships or districts.
- 2 See <https://www.rchss.sinica.edu.tw/PHD/main.php>. Researchers can apply for access to the database through the PHD website.

This paper first briefly introduces the historical institutions and contents of the Taiwanese household registers from the Japanese colonial period. Then, it offers an overview of the THHRD and its limitations. Finally, it reviews studies that made use of the THHRD and discusses the potential value of the THHRD for research.

2 HOUSEHOLD REGISTERS IN THE JAPANESE COLONIAL PERIOD

After taking over Taiwan in 1895, the Japanese soon issued an order pertaining to the implementation of a household registration system and its related regulations in the following year (Liao, 2010). The task of conducting a household survey and registering every household and its members were carried out by the military police. The registers compiled by the military police were no longer used, and were replaced by household registers of a new format after new registration regulations were promulgated on December 26, 1905. The new registers were officially established on January 15, 1906.

Instead of the military police, the new law made the municipal police and the so-called *Baochia* 保甲 in charge of the administration of household registration. The word *Baochia* denotes the head of *Bao* and the head of *Chia*. One *Chia* consisted of ten households; one *Bao* consisted of ten *Chias* (Lin, 2011; Wolf & Huang, 1980). Individuals had to report changes in their situation, including births, deaths, marriages and changes of address, to the *BaoChia* 保甲, which were the heads of the local *HoKou* 戶口 ('household') networks, within ten days, and then the police paid a home visit to verify the report (Engelen & Hsieh, 2007; Katz & Chiu, 2006). The *Baochia* functioned as a Chinese community policing system (Katz & Chiu, 2006).

To establish the initial set of the new household registers, the Japanese police reexamined the information from the defunct household registers. They interviewed family members and consulted private sources such as ancestral tablets and clan genealogies when the information appeared insufficient or missing in the records (Wolf, forthcoming). The household registration system required everyone to be registered in one household. The police also made regular door-to-door checks and irregular household surveys to make sure the accuracy of the information in the registers (Hong, 2013).

The meaning of the columns in the household registers are shown in Figure 2. If a household consisted of more than five persons, there were more pages for their household. In the Chinese household formation system, it is possible to have more than one family in a household/a house address. There were three types of household registers in colonial Taiwan: the active register, the inactive register, and finally the sojourner register. A register was a set of sheets. At the beginning of a household, the household register belonged to the active register file, when the household closed, the household registers would be moved to the inactive register file. Sojourners were the transient residents in the household. They were not part of the family and sojourned in the household because of work or some other reason. The three types of household registers shared the same registration form (Lin, 2011).

To ensure that the data in the Taiwanese household registers were accurate, they were updated regularly. Those individuals that were distrusted (vagrants and criminals) had to report to the registry office every month. For people belonging to higher social classes, this obligation was extended to every three or six months (Wolf & Huang, 1980).

The household register should be read from right to left and from top to bottom. The first item, column I, is filled in for every household register and includes the address of the household and the date and the reason for its establishment (Lin, 2011). The other columns, II to VI in Figure 2, record personal events, including the name of the other persons involved, the date when the event took place, the related address and the type of event. Sometimes a member of the household was struck out from the record, in such circumstances the individual's departure would be recorded in the personal event column (Wolf & Huang, 1980). In the case of more information, small pieces of paper would be pasted on top of it to provide additional writing space to record the new event (Lin, 2011). Each of the columns II–VI is assigned to one member of the household, except the column II which is always reserved to the household head.

Figure 2 *The household registration form in Colonial Taiwan*

VI						V						IV						III						II						I																				
Events						Events						Events						Events						Events						Present Address																				
4	3	2				4	3	2				4	3	2				4	3	2				4	3	2																								
7	6	5				7	6	5				7	6	5				7	6	5				7	6	5										Race	Address													
Relationship to head						Relationship to head						Relationship to head						Relationship to head						Relationship to head																										
																								Head of household																										
12	11	10	9	8		12	11	10	9	8		12	11	10	9	8		12	11	10	9	8		12	11	10	9	8																						

Explanation: 1. Date and reason for setting up the household, 2. Ethnic group, 3. Opium addiction, 4. Bound feet, 5. Classification (criminal records), 6. Disabilities, 7. Vaccinations, 8. Father's name, 9. Mother's name, 10. Detailed information of personal status and occupation, 11. Name, 12. Date of birth, 13. Same-sex sibling order.

Source: Chia-chi Lin (2011, p. 40).

Beneath each event column are six square boxes (box 2–7) for other personal information. The boxes are classified as 'ethnicity', 'opium addiction', 'bound feet', 'classification' (for criminal records), 'disabilities', and 'vaccination' (Wolf & Huang, 1980). In the space of 'ethnicity', people were distinguished in the light of ancestral origins, native language as well as biological affiliation (Hong, 2013; Wolf, forthcoming). The space of 'classification' (*Zhongbie* 種別) notes the rating of a person given by the police on a three-point scale. The police classified people into three groups according to their social status and behavior. There were three categories: class 1 (rich citizens and outstanding citizens); class 2 (general citizens); class 3 (pauper, criminals, entertainers and people with 'bad reputations') (Lin, 2011; Wolf, forthcoming).

The record in the opium addiction box was like a drug license. If an individual had 'A' 阿 (a contracted word of opium) written in this box, it means that he was allowed to smoke opium. Although opium abuse was harmful to an individual's health, it brought in substantial tax revenue for the Japanese government. The Japanese government therefore allowed the Taiwanese to smoke opium by buying it at official shops (Lin, 2011).

The vaccination box indicated whether an individual had received a smallpox vaccination. The box concerning 'disabilities' stated whether the person was blind, deaf or dumb. As for the record of bound feet, there were three types: 'not bound' for women who never had bound feet; 'bound' for those who had; and 'unbound' for women who had had bound feet but released them, probably because foot binding was forbidden by law in 1915. The practice of documenting criminal records (classification) was abolished in the late colonial period and the original records were covered with black ink (Lin, 2011).

As said, the horizontal rectangular box below the six square boxes indicates the relationship of each individual to the head of the household. The listing of the members was prioritized according to their relationship to the head of the household when the household register was first established. This rule was broken when somebody was married or moved into the household. In any circumstances, the first person in the record always was the head of the household (Lin, 2011).

The information collected in an individual member column (in the boxes 8–13) contains the name of a household member, the person's birthdate, his/her parents' names, his/her same-sex sibling order, and his/her relationship to the household head. The occupation of a household head and in some cases, other household members, when they had occupations which differed from the occupation of the household heads.

In 1935, the regulation of the Taiwanese household registration system was altered (Engelen & Hsieh, 2007). The information about ethnicity, police rating, physical deformities, smallpox immunization, opium smoking and foot-binding were no longer required in household registers according to the amendment (Hsu, 2014; Lin, 2011). With an exception for the occupation of a household head and in some cases other household members when they had occupations which differed from the occupation of the household heads. The sojourner register was set up only after 1935 (Chiu, 2003). Before 1935, there was no specific sojourner register book, and the sojourners were recorded together with household members in the active register. As a consequence, some sojourners have two similar records in the household registration system: one in the sojourner's original household, the other in the temporary household.

When a household member died or exited because of adoption, marriage or household division, his name was deleted from the register. When a new member arrived by means of birth, marriage or adoption, he was added to the next open column in the register. When a household head died or retired, information concerning all household members was copied onto new forms, and the old register was assigned to a file containing all the registers retired that year. When a family³ moved from one registration district to another, all information relating to its current members was copied onto fresh forms, which were sent to the registry office responsible for the household's new dwelling place. The old register was then cancelled and filed as closed registers. However, when a family moved to another village or neighborhood within the same registration district, the relevant sheets of the register were simply passed from one registry office to another without a note 'the household was quitting' in the records held by the registry office of the community (Wolf, forthcoming).

Use of the Japanese registers continued through most of 1947, despite the change of government at the end of World War II. However, the THHRD did not include any of the data recorded in 1946 and 1947 since the change of administration may have affected the reliability of such sources (Wolf, forthcoming).

3 THE DATABASE

3.1 COVERAGE

Arthur Wolf's book (forthcoming) introduces two kinds of databases of Taiwanese household registers: the Stanford Archive and the Taiwan Historical Household Registers Database, 1906–1945 (THHRD). These two databases are based on the same household registers, but include different locations and are based on different computing systems. The database of Stanford archive only consists of household registers from the Hai-san area where Arthur Wolf did his field research for many years. For more information about the Stanford Archive, see Wolf (forthcoming). In this article, we mainly introduce THHRD which is a longitudinal historical demographic database.

The THHRD database covers the period of 1906–1945. In 1906, the household registration system, as carefully designed in the household registration regulations of 1905, was put into practice. It lasted until the nationalist party took over Taiwan from the Japanese colonial government in 1945. The time span of 40 years contains up to 4 generations (Dong, Campbell, Kurosu, Yang, & Lee, 2015). By 2019 there were 369,373 individuals, and 55,810 households in THHRD. The database was constructed from various contributions. We are grateful to the collectors (see Table 1) who kindly provided their collections of household registers with the researchers.

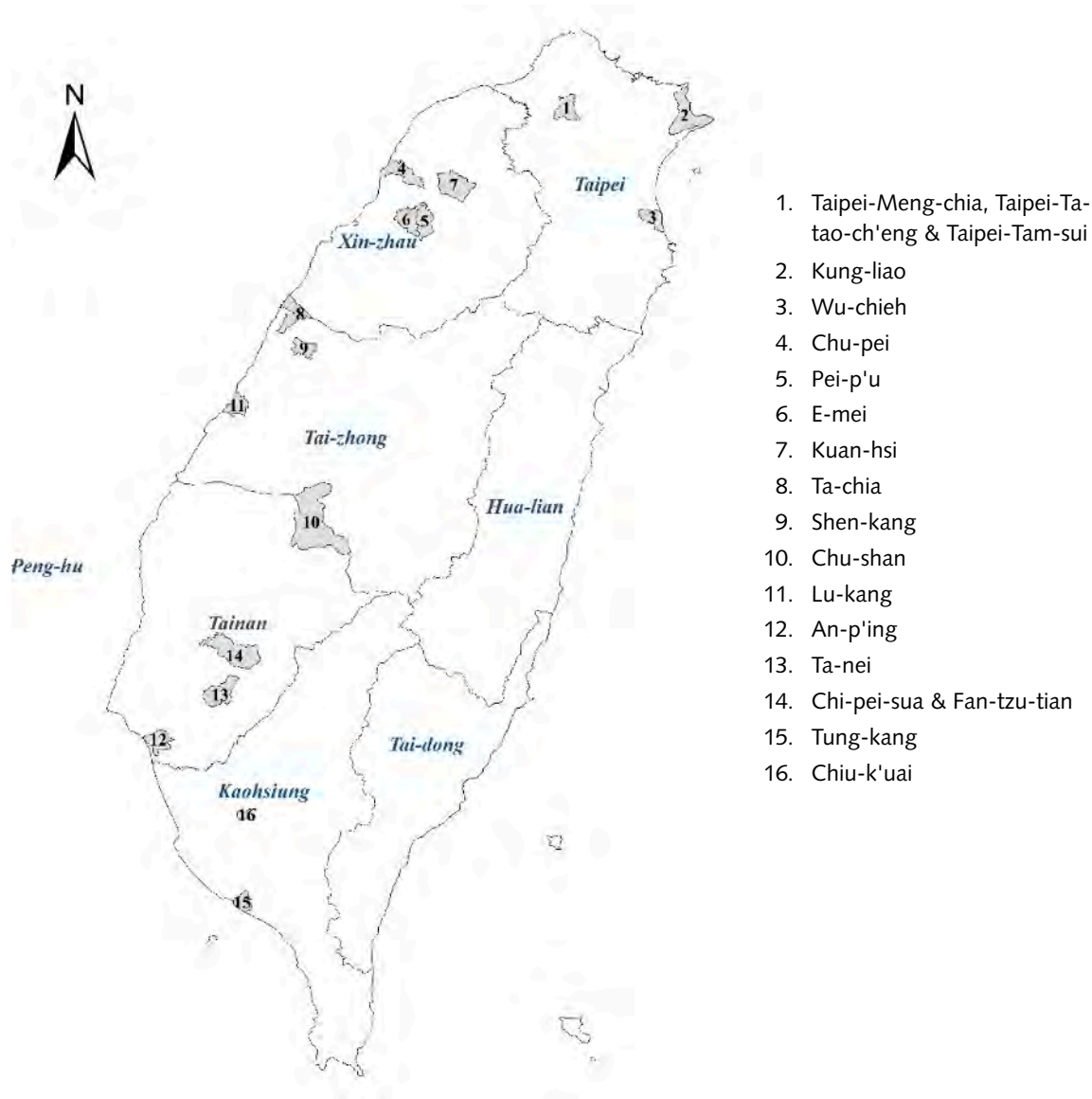
3 In traditional Chinese society, a household could include more than one family. A family consists of at least two generations.

Table 1 *Research Sites and collectors of THHRD*

Reference to figure 3	Research sites	Collectors	Numbers of individuals
1	Taipei-Meng-chia & Taipei-Ta-tao-ch'eng	Arthur Wolf, Hill Gates	35,133
	Taipei-Tam-sui	Chia-chi Lin	8,261
2	Kung-liao	Ing-hai Pan	4,066
3	Wu-chieh	Ing-hai Pan	16,780
4	Chu-pei	Ying-chang Chuang	19,671
5	Pei-p'u	Ying-chang Chuang	40,680
6	E-Mei	Ying-chang Chuang	14,216
7	Kuan-hsi	Ying-chang Chuang	53,807
8	Ta-chia	Ying-chang Chuang	11,561
9	Shen-kang	Ying-chang Chuang, Ing-hai Pan	13,854
10	Chu-shan	Ying-chang Chuang	13,241
11	Lu-kang	Guang-hong Yu	10,518
12	An-p'ing	Ing-hai Pan	17,078
13	Ta-nei	Ing-hai Pan	20,909
14	Chi-pei-sua & Fan-tzu-tian	Ing-hai Pan	4,011
15	Tung-kang	Paul Katz	11,444
16	Chiu-k'uai	Hsiang-shui Chen	4,628
Total			299,858
	Hu-hsi	Guang-hong Yu, James Wilkerson	under construction (7,508)
	Pai-sha	Guang-hong Yu	under construction (8,074)
	Ma-kung	Guang-hong Yu	under construction (16,225)

As for its spatial coverage, the THHRD contains digitized data from 30 research sites while still more localities in Taiwan are being added. Except Eastern Taiwan, they represent all of the Taiwan mainland and Pescadores Islands. Nowadays, among 30 research sites, only 16 sites are available for research, which are scattered in different regions of Taiwan (see Table 1 and Figure 3). They include eight sites in Northern Taiwan, four in Central Taiwan and four in Southern Taiwan. Among these sites, some are in urban areas, and others are in rural areas. Therefore, the THHRD allows for an observation of urban and rural differences. The mentioned program for historical demography only provides open access to the data of three research sites: Chu-pei, Pei-p'u and E-mei. As for the rest of research sites, please first contact with contributors and the program for historical demography.

Figure 3 The distribution of research sites in THHRD (2019)



3.2 CONTENTS

In the THHRD, following the design of Professor Wolf and Professor Chuang, there are 9 different tables (see Table 2) to transcribe the enormous and complex details from the household registers. The history of a household could be recorded in more than one register if the household ever experienced the death or retirement of a household head, or if one of its members ever departed to create a new household. As for an individual, he/she could also be observed in different registers in any of the previously mentioned cases as well as if he/she ever departed from his/her original household through adoption, marriage or other means.

Through linkage, household and individual histories can be reconstructed and various research tasks can be conducted. For instance, with the help of the recorded relationship of each individual to his household head, it is possible to identify the relationship between any two individuals within one household and thus to reconstruct the exact structure of a household (Wolf, forthcoming). By linking the Personal Stat Table, the Occupation Table and so on, Chia-chi Lin (2011) has published a book on female-headed households in Eurasian societies.

Table 2 *Nine information tables from household registers*

Individual-level	Household-level	Others
Personal Stat	House Stat	Address
Personal Dynamic	House Dynamic	Occupation
Personal Location		Relationship to the household head
		Time of entering or leaving

3.3 DATA LIMITATIONS

There is always a limitation in everything. As for using THHRD, users have to keep the following things in mind. First, events occurring before 1906 were usually recorded a long period after they actually happened and were mainly based on a reporter's recollection. In addition, as mentioned before, the dates of these events might have been reported according to the lunar calendar instead of the solar calendar used by the Japanese government. The lunar calendar was used in the pre-colonial farm society. This changed under the Japanese governance. So, one has to consider the accuracy of the dating of the earlier events (Hsu, 2014; Wolf, forthcoming).

Second, the way by which 'persons in temporary stay' (*Jiliouren* 寄留人) are dealt with in the database deserves some attention. Because the types of 'person in temporary stay' are of great diversity, it is hard to handle them with a computer program. Therefore, they are excluded from the Personal Dynamic Table and the Personal Location Table, while they are still included in the Personal Stat Table. The reason to include them in the Personal Stat Table is that some people in temporary stay could have blood ties with members of the household in which they were registered.

Third, information about ethnicity, police rating, physical deformities, smallpox immunization, opium smoking and foot-binding was no longer recorded in the registers after 1935.

Fourth, the household registers did not record everyone's occupations. Only 10% of the recorded occupations were from other persons in the household other than that of the head of the household. However, these records did not include the dates when the occupations were valid, but they did note changes of occupation.

Fifth, the THHRD database is based on regions. Once individuals moved out of the selected regions, there was no information in the database anymore. If individuals arrived from other regions, only the reason of moving and the address of the former household from the foregoing period were recorded.

4 POSSIBILITIES FOR RESEARCH

The THHRD has been applied to various research topics intending to understand the family composition and social structure of Taiwan in the early 20th century. The topics include adoption, ethnicity, fertility, gender differences, marriage, mortality and the related life course researches. Chuang and Wolf (1995) claim that different marriage forms have different issues of concern. Moreover, in 1996, on the cooperation of Professors Arthur P. Wolf, Ying-chang Chuang and Theo Engelen of Radboud University, the project of the 'Population and Society in Taiwan and the Netherlands' was launched. The THHRD, together with the database of the Historical Sample of the Netherlands (HSN), was utilized to make Eurasian comparative studies, and a series, *Life at Extremes*, produced four books from the project. The first book, *Marriage and the Family in Eurasia: Perspectives on the Hajnal Hypothesis*, reviewed the contribution of Hajnal's hypothesis to historical demography (Engelen & Wolf, 2005).

The second book, *Positive or Preventive? Reproduction in Taiwan and the Netherlands, 1850–1940* (Chuang, Engelen, & Wolf, 2006), discussed the Malthus hypothesis proved by the results of the Eurasia historical population database. Moreover, *Two cities, One Life. Marriage and Fertility in Lugang and Nijmegen* provided a case study furthering discussion on Eurasia difference (Engelen & Hsieh, 2007). After the marriage issue, *Death at the Opposite Ends of the Eurasian Continent: Mortality Trends in*

Taiwan and the Netherlands 1850–1945 gave information on mortality topics. Another important element in Malthus' hypothesis (Engelen, Shephard, & Yang, 2011).

The occupation titles recorded in the household registers are combined with those from the 1915 Taiwan household census into a database by Professor Chia-chi Lin. She further linked these Taiwan's historical occupation titles to HISCO (Historical International Standard Classification of Occupations) to establish 'Formosa Historical International Standard Classification of Occupations' (Formosa HISCO; see asiahisco.hisotry.tku.edu.tw), which aims to provide 'economic variables' for Taiwan's historical demography.

Professor Ko-hua Yap published a paper about foot-binding studies. In general, the Hoklo people preferred bound feet while the Hakka people preferred natural feet. He found that Hoklo female with cross-ethnic family background were more likely to get rid of foot binding, according to the household registers. Furthermore, the Hoklo women without cross-ethnic family background but close to the Hakka areas also had unbound feet as compared to their counterparts in other Hoklo areas. Professor Ko-hua Yap explained the above phenomenon as 'arms race'. That is, the more popular foot binding was in the surroundings, the more likely parents bound their daughters' feet, to avoid having a disadvantage in the marriage market. Conversely, the existence of people who did not desire bound feet in the surroundings, drastically reduced the pressure of pursuing foot binding (Yap, 2017).

Professor James Wilkerson (2010) published a chapter in a Ruizhi Lian and Ying-chang Chuang edited volume on late Qing dynasty era literati marriage. The literati status in Hsinchu at that time was a legal status for men who successfully passed one or more of the various levels of the Qing imperial exams. Both before and after marriage, women in literati families were under constant surveillance through standard Chinese style sub-bureaucracies (*BaoChia*) to impose a variety of prescriptive standards for female 'domestic decorum' (*Fudao*). When these literati are compared with non-literati, the literati had extremely low rates of violations of female domestic decorum in comparison with the high rates of violations of female domestic decorum as reported and discussed by Arthur P. Wolf and others for Hsinchu and elsewhere in Taiwan. This chapter is a series of publications discussing marriage and sub-bureaucracies in Taiwan and China.

In 2015, Professor Wen-shan Yang further joined the East Asian Population and Family History Project (EAP 2), therefore the THHRD was included in EAP 2. This project focuses on neighboring populations in East Asia that are more similar in terms of background and context. Instead of comparing European and Asian populations, it aims to point out the similarities and differences in East Asian population behavior through comparative analysis of 5 population register databases from East Asia (Dong, Campbell, Kurosu, Yang, & Lee, 2015).

5 CONCLUSION

The Taiwan Historical Household Registers Database has been known as a treasure for historical demography of Asian study. It has been established for more than 30 years and keeps growing. It has provided fruitful and successful research results. With the advancement of statistical methods and computer science, we believe there are still many issues that researchers might study by using the THHRD. This paper is a brief introduction for those who are interested in the Taiwan historical household registers database, and it is also an invitation to join us!

ACKNOWLEDGEMENTS

The source data from the Taiwan household registers used in this paper are collected by Arthur Wolf, Hill Gates, Ying-chang Chuang, Wen-shan Yang, Guang-hong Yu, Ing-hai Pan, Hsiang-shui Chen, Paul Katz, James Wilkerson, and Chia-chi Lin. Those registers were digitized as 'Taiwan Historical Household Registers Database, 1906–1945 (THHRD)' by the Program of Historical Demography,

Research Center for Humanities and Social Sciences, Academia Sinica, and are at the center of this paper. We are grateful to the members of the Program of Historical Demography for their assistance, comments, corrections, and other feedback. The analysis and results expressed here are solely those of the authors and do not necessarily represent the views of THHRD.

REFERENCES

- Chiu, C. (2003). Jihchih shinchi huchi izuliao te shinliao tase yu liyung-yl hsilaian shinchian yenchiu weili [The characteristic and using of the historiography of household registers in colonial Taiwan: A case study of Hsilaiian event]. *Taiwan shin liao yen chiu* [Journal of Taiwan historiography], 20, 94–117.
- Chuang, Y., & Wolf, A. P. (1995). Marriage in Taiwan, 1881–1905. An example of regional diversity. *The Journal of Asian Studies*, 54(3), 781–795. doi: [10.2307/2059451](https://doi.org/10.2307/2059451)
- Chuang, Y., Engelen, T., & Wolf, A. P. (Eds.). (2006). *Positive or preventive? Reproduction in Taiwan and the Netherlands, 1850–1940*. Amsterdam: Aksant Academic Publishers.
- Dong, H., Campbell, C. D., Kurosu, S., Yang, W., & Lee, J. Z. (2015). New sources for comparative social science: Historical population panel data from East Asia. *Demography*, 52(3), 1061–1088. doi: [10.1007/s13524-015-0397-y](https://doi.org/10.1007/s13524-015-0397-y)
- Engelen, T., & Hsieh, Y. (2007). *Two cities, one life: Marriage and fertility in Lugang and Nijmegen*. Amsterdam: Aksant Academic Publishers.
- Engelen, T., & Wolf, A. P. (Eds.). (2005). *Marriage and the family in Eurasia: Perspectives on the Hajnal hypothesis*. Amsterdam: Aksant Academic Publishers.
- Engelen, T., Shephard, J., & Yang, W. (Eds.). (2011). *Death at the opposite ends of the Eurasian continent: Mortality trends in Taiwan and the Netherlands, 1850–1945*. Amsterdam: Aksant Academic Publishers.
- Hong, R. (2013). *Rizhi shiqi huji dengji falu ji yongyu bianyi* [The law of household registration during the Japanese colonial period and the translation of terms]. Taichung: Household register office.
- Hsu, M. (2014). *Rizhi shiqi ji guangfu chuqi taiwan huzheng gaikuang* [An overview of household registration in Taiwan during the Japanese colonial period and the Early Recovery]. Tainan: Household register office.
- Katz, P. R., & Chiu, C. (2006). Quantifying the colonized. The history and significance of demographic sources from colonial Taiwan. In Y. Chuang, T. Engelen & A. P. Wolf (Eds.). *Positive or preventive? Reproduction in Taiwan and the Netherlands, 1850–1940* (pp. 19–38). Amsterdam: Aksant Academic Publishers.
- Liao, Y. (2010). Rizhi shiqi huji dangan zhi jianli yu yingyong: Yi yilan diqu keja yimin yanjiu weili [The establishment and application of household registration archives during Japanese colonial period: A case of Yilan county's Hakka household registers]. *Archives Semiannual*, 9(1), 40–53. Available from <https://www.airitilibrary.com/Publication/alDetailedMesh?docid=P20190425001-201003-201904250026-201904250026-40-53>
- Lin, C. (2011). *Female heads of households in eurasian societies. Taipei and Rotterdam in times of industrialization*. Taipei: Program for Historical Demography, RCHSS, Academia Sinica.
- Program for Historical Demography. (n.d.). Retrieved from <https://www.rchss.sinica.edu.tw/PHD/main.php>. Accessed on July 27, 2020.
- Wilkerson, J. (2010). Diguo, wenren yu hunyin: Qingmuo Zhujian [Hsinchu] wenren jiating zhong di nuxing hunying xingshi chutan [Empire, literati, and marriage: A preliminary discussion of female marital patterns in Zhujian [Hsinchu] literati families late in the Qing Dynasty]. In R. Lian & Y. Chuang (Eds.). *Hakka, women, and marginality* (pp. 161–206). Taipei: Nantian shuju.
- Wolf, A. P. (1968). Adopt a daughter-in-law, marry a sister: A Chinese solution to the problem of the incest taboo. *American Anthropologist*, 70(5), 864–874. doi: [10.1525/aa.1968.70.5.02a00040](https://doi.org/10.1525/aa.1968.70.5.02a00040)
- Wolf, A. P. (forthcoming). *Records of a natural experiment*. Taipei: Program for Historical Demography, RCHSS, Academia Sinica.
- Wolf, A. P., & Huang, C. (1980). *Marriage and adoption in China, 1845–1945*. Stanford: Stanford University Press.
- Yap, K. (2017). Dang chanzu yushang tianranzu: Zuqun ronghe yu shehui yali [When bound feet encounter natural feet: Ethnic assimilation and social pressure]. *Minsu Quyi* [Journal of Chinese Ritual, Theatre and Folklore], 197, 107–133. doi: [10.30157/JCRTF](https://doi.org/10.30157/JCRTF)

The 2020 IDS Release of the Antwerp COR* - Database

Evaluation, Development and Transformation of a Pre-Existing Database

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

ABSTRACT

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

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

DOI article: <https://doi.org/10.51964/hlcs9301>

© 2020, Jenkinson, Anguita, Paiva, Matsuo, Matthijs

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

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

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

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

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

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

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

2 THE ANTWERP COR*-DATABASE

2.1 SAMPLING METHOD

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

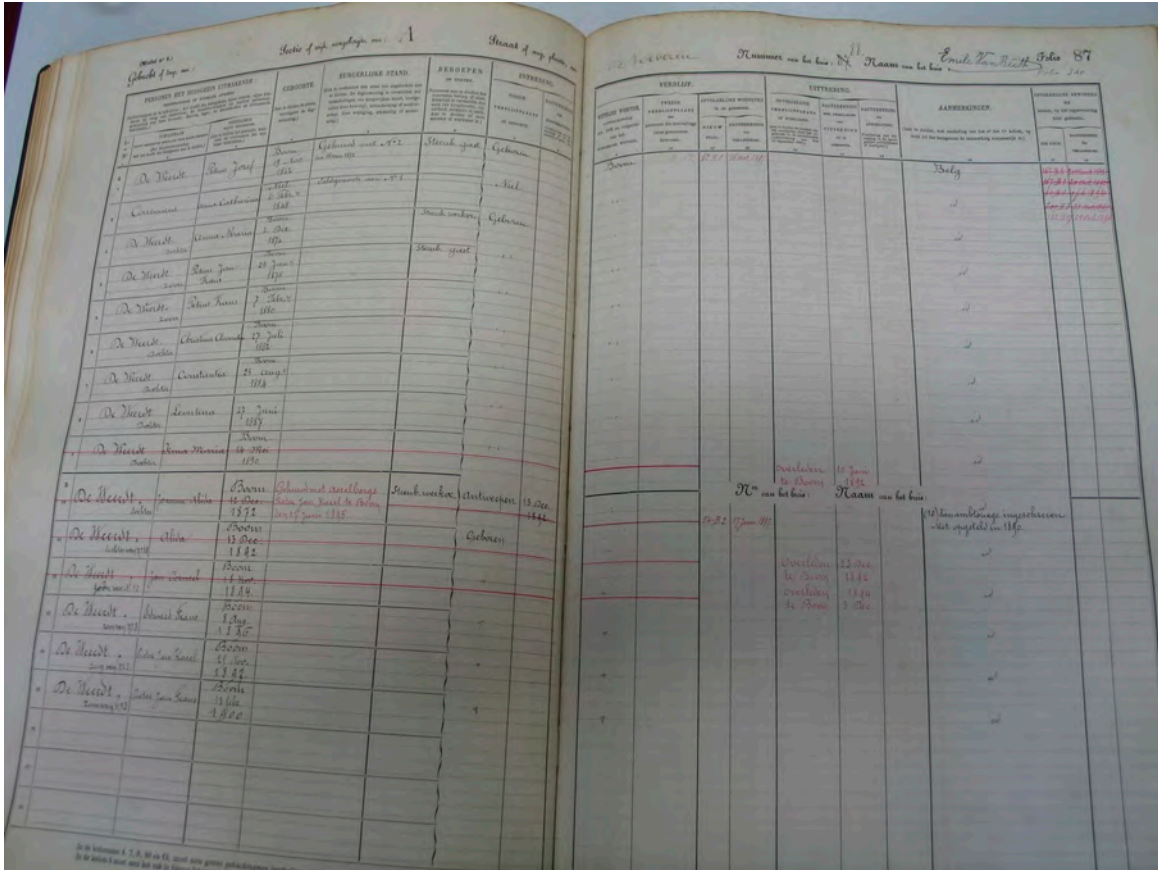
The choice of family names starting with the letters 'COR' had a number of motivations aimed at strengthening the representativeness of the database, and the choice was made in consultation with the Faculty of Linguistics at the University of Leuven. These names were shown to be evenly geographically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

2.2 ORIGINAL SOURCES

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

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

Figure 1 Picture of a Belgian population register



3 DATABASE DEVELOPMENTS

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

3.1 SAMPLE EVALUATION

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

We began by splitting the database into datasets of the original source records (i.e. source of data). Following this we proceeded by comparing the identifying attributes belonging to an IDNR across sources in order to test for discrepancies. Concerning death dates, the number of discrepancies across records for the same individual was below 1% for day, month and year examined separately. For births this was also

the case, with the exception of year, where 3% were inaccurate. Regarding birth and death locations, the discrepancies were much higher, however this is largely due to differing spellings of place names, as well as stray capitalisation and white space, which took place during the data entry process. This is standardised within the database. For gender the number of discrepancies was 0.83%. We believe the low number of discrepancies is supportive of the high quality of the original record linkage.

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

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

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

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

In order to evaluate this parental linkage contained within the 2010 release of the Antwerp COR*-database, the first step taken was to compare parental information by individuals between different sources. For this

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

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

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

3.2 RECONSTRUCTION OF RELATIONSHIPS WITHIN FAMILIES

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

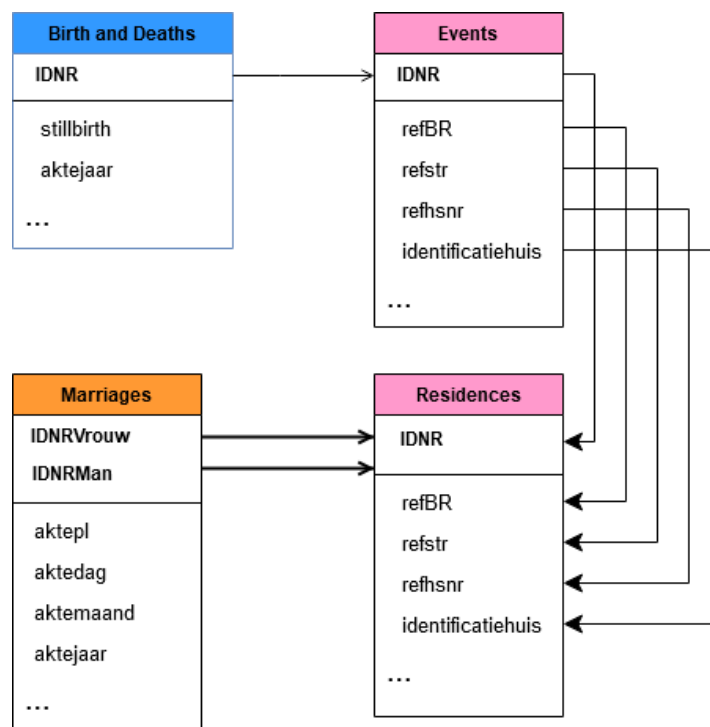
3.2.1 OVERVIEW AND INPUTS

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

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

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

Figure 2 2010 release of the Antwerp COR*-database



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

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

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

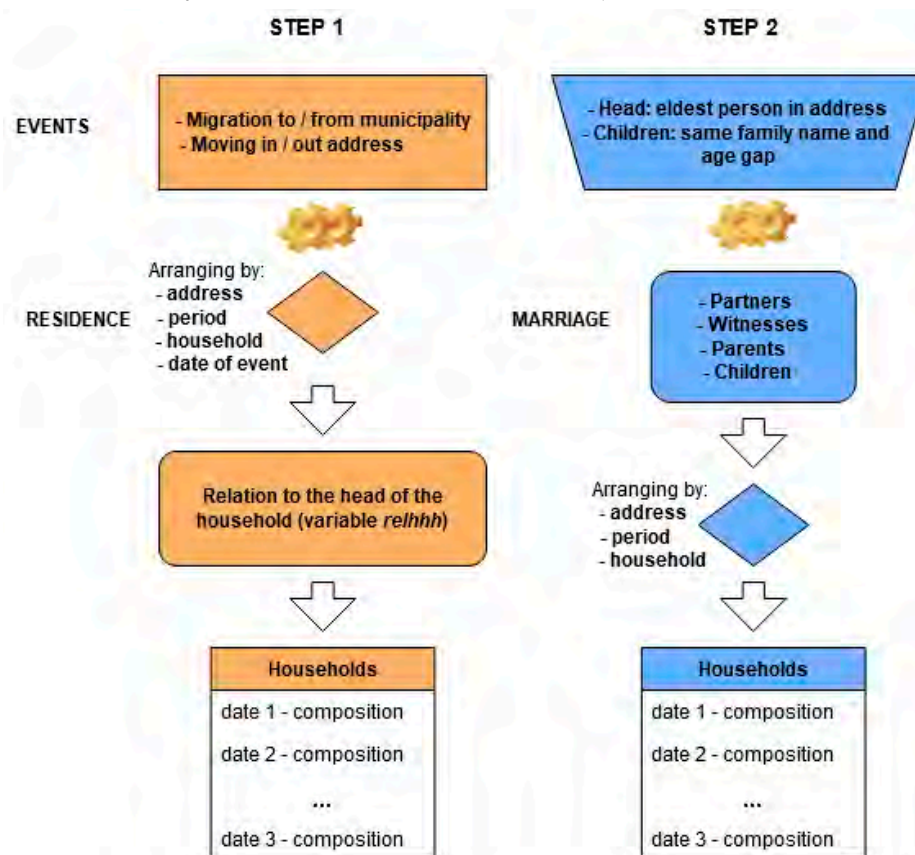
3.2.2 METHODOLOGICAL STEPS FOR THE RECONSTITUTION OF FAMILIAL RELATIONSHIPS

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

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

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

Figure 3 Visual representation of reconstitution of family method



In the first step of Figure 3, we focus on the events of moves, and this information is merged with that contained in the *Residence* tables. Its outcome is arranged by each record's identification (or IDNR), the address, the time period of the address registration in the official municipality records, and the pre-defined identification number of the household. Through examining events involving moves, we were able to

identify conjoint actions of different persons.¹ This allowed us to reconstruct some of the relations between individuals sharing family links, by creating groups or clusters of people sharing an address at a particular time. For this goal, the utilization of the information of the role of individuals with respect to the head person was also useful (when this information was available). This is collected by the variable 'relhjh' that we can see in Figure 3. Through the combination of all these actions we managed to reconstruct the first version of the address based clusters (see the final stage of Step 1 in Figure 3, outlining their prospective composition throughout time).

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

3.2.3 HOUSEHOLDS AND HEADS

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

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

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

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

3.2.4 PARENTS AND CHILDREN

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

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

3.3 GEOCODING THE ANTWERP COR*-DATABASE

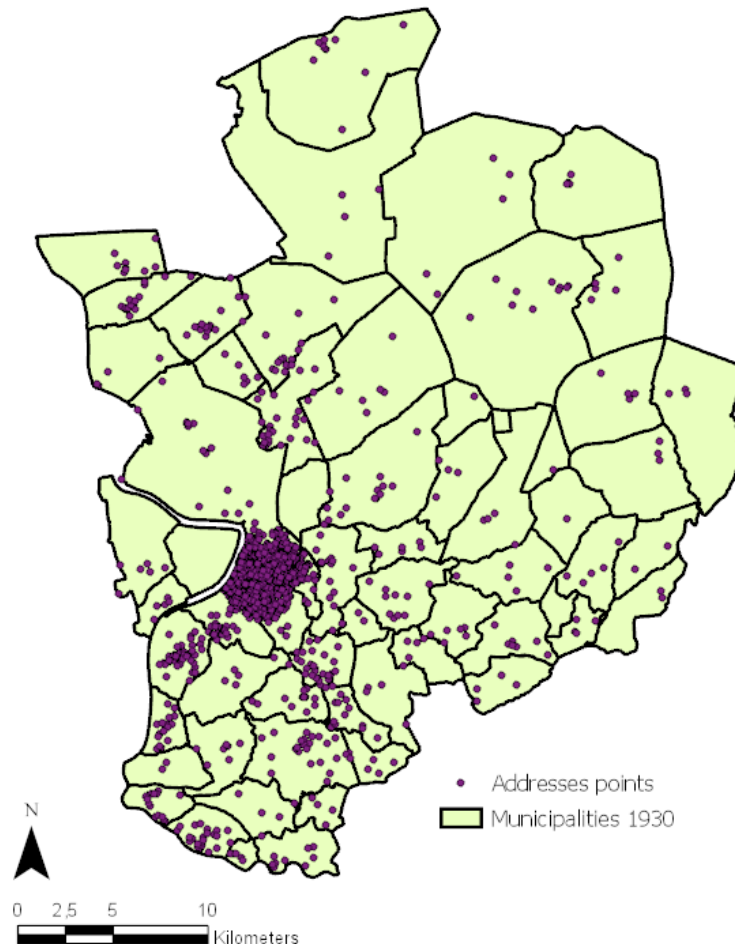
Spatial analysis is continuing to become an ever more important analytical component to an ever increasing number of highly significant demographic research questions. We follow the similar work by Hedefalk, Harrie, and Svensson (2014) (IDS-Geo), namely the inclusion of a coordinate system for the location of address level information. This development adds critical value to the current format of the 2010 release of the Antwerp COR*-database, by specific means of additional variables: a twofold system in the context table (street centroid² and municipality centroid³) and including individual variables at the very low level (i.e. street level) (i.e. NIS level 7⁴) (Paiva, 2019). Figure 4 represents the municipalities in the Antwerp region, while figure 5 shows the representation of addresses by street centroids.

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



2 This means for instance, the median point of the line representation of the street.
 3 A point coordinate representing the location of the administrative-political centre of the municipality.
 4 NIS code represents an alphanumeric code for regional areas and consists of 5 digits. The first number refers to the province; the second, the arrondissement within the province and the last three, identifies the community within the arrondissement (Statistics Belgium).

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



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

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

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

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

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

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

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

Figure 6 Oorderen area 19th century



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

- 6 For example: 'Driesch', 'Driessche', 'Drieschstraet' and 'Dries Straat' in Antwerp were all being standardized as 'Driesstraat' while the correct spelling is 'Dries'; more significantly, 'Hagelkruisstraat', 'Hagelkruis' (both as 'Hagelruis'), 'Gr. Hagelkruisstraat' (as 'Klein Hagelkruis') were being divided into two different streets although all are forms of 'Groot Hagelkruis' and the latter was transformed from Groot ('Gr.') to Klein. The process of standardization relied on the document of 'alle plaatsen' recorded for the production of COR*2010.
- 7 GISTorical Antwerp project provided an extensive list of old street names (mostly for Antwerp and its surroundings: Berchem, Hoboken, Borgerhout and Oosterweel), Robert vande Weghe's *Geschiedenis van de Antwerpse straatnamen* (1977).

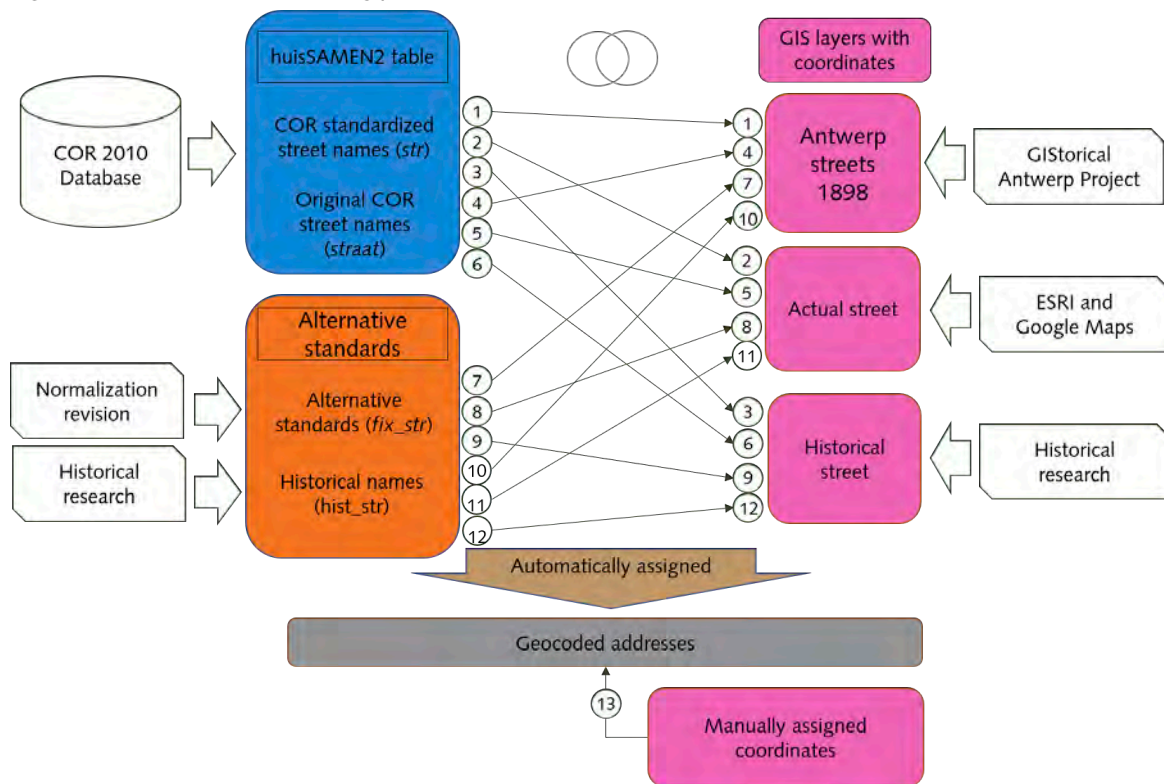
Figure 7 Oorderen area today



Source: www.geopunt.be.

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

Figure 8 The Geocoding process



3.4 2020 IDS RELEASE OF THE ANTWERP COR*-DATABASE

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

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

3.4.1 IMPERFECT DATA AND TIME STAMPS

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

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

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

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

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

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

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

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

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

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

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

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

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

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

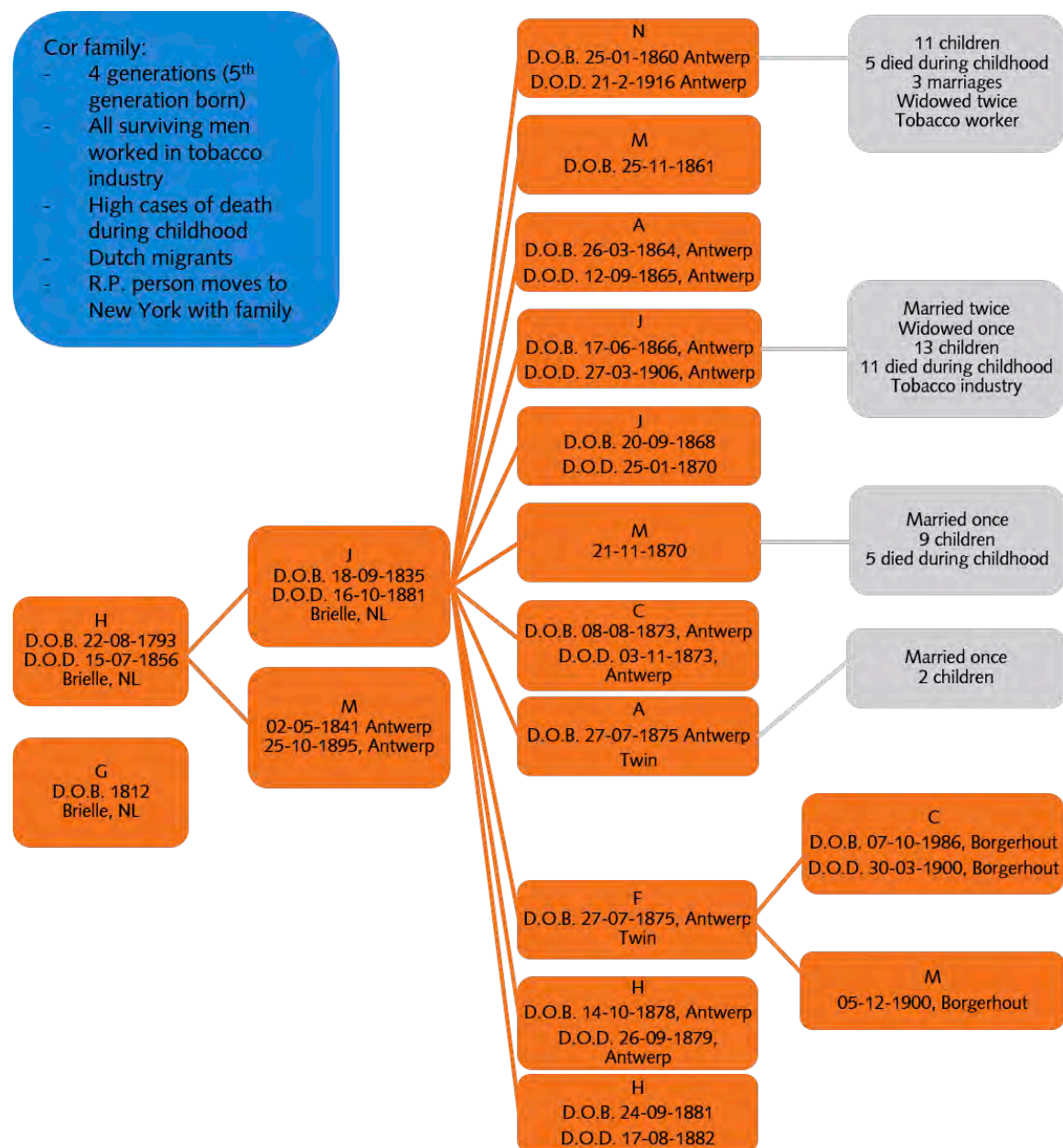
A further unique characteristic of the Antwerp COR*-database concerns occupational level information, recorded from population registers and also vital registration records (i.e. parental information from birth, marriage and death certificates). Depending on the individuals, it is quite often possible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

3.4.4 MICRO EXAMPLE: LIFE HISTORIES IN THE INTERGENERATIONAL HOUSEHOLD

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

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

Figure 9 *Life histories across generations*



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

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

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

4 DISCUSSION AND CONCLUSION

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

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

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

this new asset to the 2020 IDS release of the Antwerp COR*-database would allow similar insightful research for the Antwerp arrondissement.

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

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

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

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

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

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

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

ACKNOWLEDGEMENTS

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

REFERENCES

- Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal microdata, version 4. *Historical Life Course Studies*, 1(1), 1–26. Retrieved from <http://hdl.handle.net/10622/23526343-2014-0001?locatt=view:master>
- Bourdieu, J., Kesztenbaum, L., Postel-Vinay, G. (2014). *L'enquête TRA, histoire d'un outil, outil pour l'histoire: Tome 1 (1793–1902)*. Paris, INED: Classiques de l'économie et de la population

- Broström, G., Edvinsson, S., & Engberg, E. (2018). Intergenerational transfers of infant mortality in 19th-century northern Sweden. *Historical Life Course Studies*, 7(2), 106–122. Retrieved from <http://hdl.handle.net/10622/23526343-2018-0005?locatt=view:master>
- Contiero, P., Tittarelli, A., Tabliabue, G., Maghini, A., Fabino, S., Crosignan, P., & Tessandori, R. (2005). The Epilink record linkage software presentation and results of linkage test on cancer registry files. *Methods of Information in Medicine*, 44(1), 66–71.
- Donrovich, R., Puschmann, P., & Matthijs, M. (2018). Mortality clustering in the family. Fast life history trajectories and the intergenerational transfer of infant death in late 19th- and early 20th-century Antwerp. *Historical Life Course Studies*, 7, 47–68. Retrieved from <http://hdl.handle.net/10622/23526343-2018-0006?locatt=view:master>
- Ekamper, P., & van Poppel, F. W. A. (2019). Infant mortality in mid-19th century Amsterdam: Religion, social class, and space. *Population, Space and Place*, 25(4). doi: 10.1002/psp.2232
- Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183–1210. doi: 10.1080/01621459.1969.10501049
- Geopunt.be. (n.d.). *Atlas der Buurtwegen (1841). Vandermaelen kaarten (1846–1854) and Popp kaarten (1842–1879)*. <https://www.geopunt.be/>. Accessed at June 27, 2020.
- Haber, M. (1984). Algorithm AS 207: Fitting a general log-linear model. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(3), 358–362. doi: 10.2307/2347724
- Hedefalk, F., Harrie, L., & Svensson, P. (2014). Extending the intermediate data structure (IDS) for longitudinal historical databases to include geographic data. *Historical Life course studies*, 1, 27–46. Retrieved from <http://hdl.handle.net/10622/23526343-2014-0003?locatt=view:master>
- Jenkinson, S., Matsuo, H., & Matthijs, K. (2017). *COR technical note for the construction of intermediate data structure (IDS)*. (LONGPOP research output 7.1.). LONGPOP. Retrieved from http://longpop-itn.eu/wp-content/uploads/2018/05/S.Jenkinson_COR_technical_note_construction_IDS.pdf
- Jenkinson, S., Matthijs, K., & Matsuo, H. (2019). *COR*-IDS progress report: Inter-generational linkage*. (LONGPOP research output 7.3/4). LONGPOP. Retrieved from https://limo.libis.be/primo-explore/fulldisplay?docid=LIRIAS2818367&context=L&vid=Lirias&search_scope=Lirias&tab=default_tab&lang=en_US&fromSitemap=1
- LONGPOP (n.d.). Website: <http://longpop-itn.eu/>.
- Loyen, R. (2003). Throughout in the port of Antwerp (1901–2000): An integrated functional approach. In R. Loyen, E. Buyst & G. Devos (Eds.), *Struggling for leadership: Antwerp-Rotterdam Port Competition between 1870–2000* (pp. 29–61). Heidelberg, Germany: Physica-Verlag.
- Matthijs, K., & Moreels, S. (2010). The Antwerp COR*-database: A unique Flemish source for historical-demographic research. *The history of the family*, 15(1), 109–115. doi: 10.1016/j.hisfam.2010.01.002
- De Mulder, W., & Neyrink, W. (2014). *Documentation construction IDS database with Antwerp COR*-data*. (WOG report Historical Demography WOG/HD/2014-1). Leuven: CeSO. Retrieved from https://soc.kuleuven.be/ceso/fapos/nasdltrc/files/WOG2014-1OmzettingCORnaarIDS_09012015.pdf
- Paiva, D. (2019). *Geocoding COR*-Antwerpen Database*. (LONGPOP research output). Retrieved from http://longpop-itn.eu/wp-content/uploads/2019/07/D.Paiva_Geocoding_COR-2_database.pdf
- Puschmann, P., Gröberg, P.-O., Schumacher, R., & Matthijs, K. (2014). Access to marriage and reproduction among migrants in Antwerp and Stockholm. A longitudinal approach to processes of social inclusion and exclusion 1846–1926. *The History of the Family*, 19(1), 29–52. doi: 10.1080/1081602X.2013.796889
- Puschmann, P. (2015). *Social inclusion and exclusion of urban in-migrants in northwestern European port cities. Antwerp, Rotterdam & Stockholm ca. 1850–1930*. (PhD thesis, KU Leuven). Retrieved from https://www.researchgate.net/publication/287106072_Social_Inclusion_and_Exclusion_of_Urban_In-Migrants_in_Northwestern_European_Port_Cities_Antwerp_Rotterdam_Stockholm_ca_1850-1930
- Quaranta, L. (2018). Program for studying intergenerational transmissions in infant mortality using the Intermediate Data Structure (IDS). *Historical Life Course Studies*, 7, 11–27. Retrieved from <http://hdl.handle.net/10622/23526343-2018-0010?locatt=view:master>

- Quaranta, L., Broström, B., van Dijk, I., Donrovich, R., Edvinsson, S., Engberg, E., Mandemakers, K., Matthijs, K., Puschmann, P., & Sommerseth, H. L. (2017, April 27). Intergenerational transfers of infant mortality in historical contexts: a comparative study of five European populations. Paper presented at the Population Association of America, Chicago. Retrieved from <https://paa.confex.com/paa/2017/meetingapp.cgi/Paper/15094>
- Sariyar, M., & Borg, A. (2010). The record linkage package: Detecting errors in data. *The R Journal*, 2(2), 61–67. Retrieved from <https://journal.r-project.org/archive/2010/RJ-2010-017/index.html>
- Sariyar, M., & Borg, A. (2016). *R package 'record linkage'*. Package retrieved from <https://cran.r-project.org/web/packages/RecordLinkage/index.html>
- Sommerseth, H. L. (2018). The intergenerational transfer of infant mortality in Northern Norway during the 19th and early 20th centuries. *Historical Life Course Studies*, 7, 69–87. Retrieved from <http://hdl.handle.net/10622/23526343-2018-0008?locatt=view:master>
- Statistics Belgium. (2020). *Geografische indeling (geographical information)*. Retrieved from <https://statbel.fgov.be/nl/over-statbel/methodologie/classificaties/geografie>. Accessed on September 4, 2020.
- Van Baelen, H. (2007). *Constructie van een historisch-demografisch longitudinale database: Methodologie van de Demographica Flandria selecta*. Leuven: CeSO.
- van Dijk, I. K., & Mandemakers, K. (2018). Like mother, like daughter. Intergenerational transmission of infant mortality clustering in Zeeland, the Netherlands, 1833–1912. *Historical Life Course Studies*, 7, 28–46. Retrieved from <http://hdl.handle.net/10622/23526343-2018-0003?locatt=view:master>
- vande Weghe, R. (1977). *Geschiedenis van de Antwerpse straatnamen*. Antwerp: Mercurius.
- Viciana, F. (2020). *viciana/RTransposer: Import data from R data.table Github*. Retrieved from <https://github.com/viciana/RTransposer>. Accessed on November, 5, 2020.
- Vrielinck, S., Wiedemann, T., & Deboosere, P. (n.d.). *HISGIS België 1800–2000*.
- Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354–369. Retrieved from <https://files.eric.ed.gov/fulltext/ED325505.pdf>
- Winter, A. (2009). *Migrants and urban change. Newcomers to Antwerp, 1760–1860*. London: Pickering & Chatto /Routledge.

APPENDIX

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

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

The Barcelona Historical Marriage Database and the Baix Llobregat Demographic Database

From Algorithms for Handwriting Recognition to Individual- Level Demographic and Socioeconomic Data

Joana Maria Pujadas-Mora	Open University of Catalonia & Center for Demographic Studies, Autonomous University of Barcelona
Alícia Fornés	Computer Vision Center, Autonomous University of Barcelona
Oriol Ramos Terrades	Computer Vision Center, Autonomous University of Barcelona
Josep Lladós	Computer Vision Center, Autonomous University of Barcelona
Jialuo Chen	Computer Vision Center, Autonomous University of Barcelona
Miquel Valls-Fígols	Center for Demographic Studies, Autonomous University of Barcelona
Anna Cabré	Center for Demographic Studies, Autonomous University of Barcelona

ABSTRACT

The Barcelona Historical Marriage Database (BHMD) gathers records of the more than 600,000 marriages celebrated in the Diocese of Barcelona and their taxation registered in Barcelona Cathedral's so-called Marriage Licences Books for the long period 1451–1905 and the BALL Demographic Database brings together the individual information recorded in the population registers, censuses and fiscal censuses of the main municipalities of the county of Baix Llobregat (Barcelona). In this ongoing collection 263,786 individual observations have been assembled, dating from the period between 1828 and 1965 by December 2020. The two databases started as part of different interdisciplinary research projects at the crossroads of Historical Demography and Computer Vision. Their construction uses artificial intelligence and computer vision methods as Handwriting Recognition to reduce the time of execution. However, its current state still requires some human intervention which explains the implemented crowdsourcing and game sourcing experiences. Moreover, knowledge graph techniques have allowed the application of advanced record linkage to link the same individuals and families across time and space. Moreover, we will discuss the main research lines using both databases developed so far in historical demography.

Keywords: Individual demographic databases, Computer vision, Record linkage, Social mobility, Inequality, Migration, Word spotting, Handwriting recognition, Local censuses, Marriage licences

DOI article: <https://doi.org/10.51964/hlcs11971>

© 2022, Pujadas-Mora, Fornés, Ramos Terrades, Lladós, Chen, Valls-Fígols, Cabré
This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

The Barcelona Historical Marriage Database (BHMD) gathers records of the more than 600,000 marriages celebrated in the Diocese of Barcelona and their taxation registered in Barcelona Cathedral's so-called Marriage Licenses Books for the long period 1451–1905. The Baix Llobregat Demographic Database (BALL) brings together the individual information recorded in the population registers, censuses and fiscal censuses of the main municipalities of the county of Baix Llobregat (Barcelona). In this ongoing collection 263,786 individual observations have been assembled, dating from the period between 1828 and 1965 by December 2020. The two databases started as part of different interdisciplinary research projects at the crossroads of Historical Demography and Computer Vision, within the well-known Digital Humanities, by a team of demographers, historians, geographers, mathematicians, and computer scientists from the Geography Department of Universitat Autònoma de Barcelona, the Center for Demographic Studies (CED), and the Computer Vision Center (CVC). To this end, these databases are conceived not only as demographic databases, but also as interdisciplinary research infrastructures.

Artificial intelligence methods are used to reduce the time needed for the construction of the database execution and to cope with the expansive volume of data, since gathering original information from primary sources is a time-consuming process, even when using a computer. In particular, we use computer vision technologies to extract information from the scanned document images. Computer Vision is the subfield of Artificial Intelligence that designs software to give machines the ability to see. In this work, Computer Vision aims to automatically read the text of original handwritten sources. The current development of optical character recognition, including the successful adoption of deep learning and machine learning in handwritten text recognition (Lladós, Rusinol, Fornés, Fernández, & Dutta, 2012; Toledo, Dey, Fornés, & Lladós, 2017), opens up the possibility of their integration into data collection to shorten the process. It also contributes to expand the volume of information in line with the big data revolution. However, the present approaches are specific for each data source considering language, structure, vocabulary, etc. Notably, these techniques are moving towards 'Document Understanding' to narrow the semantic gap between interpreting the original contents and automatically extracting and filling a database (Toledo, Carbonell, Fornes, & Lladós, 2019). Integrating these techniques brings Historical Demography into the realm of Big Data Sciences and allows the study of new topics and revisitation of old ones from a life course perspective, and for taking a multigenerational approach.

The research group responsible for constructing these databases was originally created to respond to the European Research Council's call for Advanced Grant projects, obtaining the 'Five Centuries of Marriages — 5CofM' project for the period 2011–2016 directed by Prof. Anna Cabré, one of the main objectives of which was the creation of the BHMD. This project was followed by others such as 'Tools and procedures for the large scale digitization of historical sources of population — *Eines*' (PI: Josep Lladós and Albert Esteve) (2015–2017) and 'Technology and citizen innovation for building historical social networks to understand the demographic past — *Xarxes*' (PI: Alicia Fornés and Joana Maria Pujadas-Mora) (2017–2019), both supported by the RecerCaixa program, and 'Demographic determinants of economic inequality, a historical approach (18th–20th centuries) — *Demodesigual*' funded by the Spanish Ministry of Sciences (PI: Joana Maria Pujadas-Mora) (2019–2021), under which the BALL database was built.

This article is organized in seven sections. Section 2 describes the source materials on which the BHMD and the BALL are based. In the following section, the data structures and data harmonization for the two databases are discussed. In section 4, we describe how the databases were constructed, presenting the use of web-based crowdsourcing platforms to transcribe the original sources, which were validated using game sourcing mobile applications, both assisted by computer vision techniques. In section 5, data linkage is presented to show how string distances, visual word search, and knowledge graph-based methods are useful to build individual life courses and multigenerational families. Section 6 is devoted to the research fields addressed, involving both databases, in historical demography, as the study of intergenerational transmission of social status, intermarriage and kinship marriages within the social reproduction process, the evolution of migratory flows in a long-term view and the estimation of economic inequality within preindustrial and industrial periods (15th–20th centuries). The paper closes with some final comments on the future agenda to expand the two databases with other textual sources and images.

2 SOURCE MATERIALS

2.1 MARRIAGE LICENSES BOOKS

The BHMD brings together the 612,487 marriages recorded in the so-called Marriage Licenses Books of the *lus Tabulae* of the Cathedral of Barcelona, covering the Diocese of Barcelona between 1451 and 1905. This is a unique data source dating back to 1409 when Pope Benedict XIII (1328–1423) visited Barcelona and granted the new cathedral the power to impose a tax on marriage in accordance with the socio-economic status of the couple to fund the cathedral's construction and maintenance (Baucells, 2002; Carreras Candi, 1913). The names and surnames of the grooms were registered (one surname up to 1876 and two surnames thereafter), while the names of the brides started to be registered from 1481 onwards. Previously unmarried brides were related to their fathers and widows to their late husbands. The occupations of the grooms, the marital status of both spouses, and the tax paid were also recorded throughout the entire life of the source, as were the first and surnames of the spouses' parents, except for the period 1645–1715. In fact, the availability of occupational and fiscal information for so many centuries is one of the main strengths of this database, enabling us to adopt a long-term perspective approach in our studies. However, the occupations of the fathers were not systematically recorded except for the period 1545–1643 and the occupations of the wives and mothers were never recorded. The original books were written in Catalan until 1860, after which Spanish was used (see Table 1).

From 1575 onwards, marriage taxes were organized on a seven- or eight-tiered scale), ranging from the highest tax paid by the nobility to exempt from tax for those declared poor. From 1451 to 1565, there were up to 27 different payment levels, while seven levels were set from 1583 onwards. In this seven-level system the first level corresponded to the titled nobility and the next two to the knights and honest citizens, or those who could hold public office. The fourth and fifth levels of payment corresponded to the commercial bourgeoisie, liberal professionals, and masters of guilds. The sixth was paid by farmers and artisans, and the last level was the exempt from tax category. A new level was already added in 1649 and continued till 1857, corresponding exclusively to the merchants of Barcelona. The highest tax level was 120 times higher than the lowest one and 40 times more than the average tax.

As mentioned previously, the territorial coverage of the BHMD is the Diocese of Barcelona (see Map 1), which was made up of four deanships: Barcelona, the main one, which is why it was called Officiality, Piera, Vallès and Penedès. This territory covered the main population centers of the time including Barcelona, Mataró, Sabadell, and Terrassa, and a conglomerate of rural towns located in the current counties of Baix Llobregat, Barcelonès, Maresme, and Vallès Occidental (see Map 1). In 1900, the diocese was comprised of 250 parishes. This source is not only extraordinary for its territorial coverage and chronological amplitude, starting a few centuries before the parish marriage books, but also for its state of preservation compared with the low conservation of the parish archives in Catalonia, especially in the study area.

Map 1 Geographical coverage of BHMD and BALL database. Area of the Barcelona Diocese



Table 1 *Number of marriages and their compiled information in the marriage licenses books, 1451–1905*

		Marriage Licenses								
		1450– 1499(a)	1500– 1549	1550– 1599	1600– 1649	1650– 1699(b)	1700– 1749	1750– 1799(c)	1800– 1849	1850– 1905
Number of marriages		7,002	23,766	44,567	57,050	41,718	69,744	82,608	96,579	189,453
General	Priest first name and surname	X								
	Date of the payment	X	X	X	X	X	X	X	X	X
	Parish of celebration						X	X	X	X
	Fee	X	X	X	X	X	X	X	X	X
Groom	First name	X	X	X	X	X	X	X	X	X
	Surname 1	X	X	X	X	X	X	X	X	X
	Surname 2									X
	Marital status				X	X	X	X	X	X
	Occupation	X	X	X	X	X	X	X	X	X
	Geographical origin		X	X	X	X	X			
	Residence	X	X	X	X					
Bride	First name		X	X	X	X	X	X	X	X
	Surname 1					X	X	X	X	X
	Surname 2									X
	Marital status		X	X	X	X	X	X	X	X
	Occupation									
	Geographical origin									
	Residence									
Father of the groom	First name			X	X			X	X	X
	Surname 1			X	X			X		
	Surname 2									X
	Marital status									
	Occupation			X	X					
	Geographical origin									
	Residence			X	X					
Mother of the groom	First name			X	X			X	X	X
	Surname 1							X	X	X
	Surname 2									
	Marital status									
	Occupation									
	Geographical origin									
	Residence									
Father of the bride	First name	X	X	X	X			X	X	X
	Surname 1	X	X	X	X			X		
	Surname 2									
	Marital status									
	Occupation		X	X	X					
	Geographic origin			X	X					
	Residence	X	X	X	X					
Mother of the bride	First name			X	X			X	X	X
	Surname 1					X		X	X	X
	Surname 2									
	Marital status									
	Occupation									
	Geographic origin									
	Residence									

Notes: (a) Except 1456–1480 and 1491–1499; (b) Except 1653–1659; (c) Except 1782–1784.

2.2 POPULATION REGISTERS, NATIONAL CENSUSES AND PERSONAL TAXES

The BALL database of which the construction is still ongoing, compiles the population registers — called *padrones* in Spanish — and censuses of 14 municipalities in the county of Baix Llobregat (province of Barcelona) and the personal taxes of nine of these municipalities collected between the 19th and 20th centuries (see Map 1). In Spain, the population registers were introduced with the Decree issued on 3 February 1823 being household registers. Initially, there was no simultaneity or fixed periodicity in the population registers until a five-year interval was enacted by the Municipal Law of August 20th, 1870, which was followed mainly from the beginning of the 20th century. The population registers contain information about individuals and households for urban centers, small villages or hamlets, and isolated houses. The first population registers in the 1820s and 1830s were simply lists of inhabitants, becoming complete registers from the end of the 1880s onwards, recording the first name and surnames, age or date of birth, marital status, and occupation of each individual, in addition to the family or the working relationship with the head of the household and the complete address. For some periods and municipalities, there is also information about the individuals' literacy skills, migratory status, and income (see Table 2). The variability of information among the registers of the different municipalities was because the number and type of variables were not standardized for the whole country until the Municipal Statute of March 8th, 1924 (García Pérez, 2007; García Rui Pérez, 2012). All this information may originally have been used for statistical purposes, as a way of counting the number of inhabitants or their main socioeconomic and demographic features in national censuses, but unlike the national censuses their main purpose was in fact to meet the military and fiscal requirements of the State and to provide proof of residence in the municipality.

The population registers contain almost the only census data preserved at the individual level in Spain because national census manuscripts were generally destroyed once the population count had been estimated and the main variables aggregated, as established by law. Therefore, only the census tabulations are preserved in the current National Institute of Statistics. However, the individual registrations of the national censuses were found for different municipalities in the Baix Llobregat county. Wherever the national censuses had been preserved locally they were also included in BALL even though the information they provided was very similar to that contained in the population registers. Moreover, the national censuses tended to provide more information of a socioeconomic nature or about literacy. And last, this meant an increased number of observations in the BALL mainly from the 20th century onwards because the censuses were carried out in the years ended in 0, while the population registers were compiled in the ones ending in 5 (Reher & Valero Lobo, 1995). In this way, the population registers throughout the 19th century are quite similar to the usual population registers in the north of Europe given their high cadence in time — especially in certain municipalities — although the consecutive years have not necessarily been collected in the BALL database, because of the higher priority of expanding the geographical and time coverage. On the other hand, in the 20th century they are more similar to the national censuses being snapshots in time.

The first modern national census in Spain was carried out in 1857, although there are direct precedents such as the censuses of Aranda (1768), Floridablanca (1787), and Godoy (1797). While we did not find the individual information, we did find the aggregated data. From 1857 onwards, national censuses began to take on the distinguishing features of modern population censuses sponsored by the State with appropriate administrative procedures and legislation, national boundaries, universality, individual enumeration, simultaneity in data collection, periodicity, publication, and diffusion (Goyer & Draaijer, 1992; Reher & Valero Lobo, 1995). National censuses were also taken in 1860, 1877, 1887, 1897, 1900, 1910, 1920, 1930, 1940, and so on until 2011 when the census started to be based on the population registers combined with a 10% survey. From 2021 the national census will only be constructed using administrative data as in the Netherlands, Denmark, Sweden and other European countries. Notably, when the state started to compile the national censuses in the mid-19th century, the public administration with the most experience in compiling demographic data in population registers and in which the largest number of civil servants or public employees were involved was the municipal administration (Salas-Vives & Pujadas-Mora, 2021).

In 1854, the personal taxes, first called *cédulas de vecindad*, later *cédulas de empadronamiento*, and finally, *cédulas personales*, were established as capitation taxes and were a type of personal identity document up until 1943 (Martín Niño, 1972). It is a source that is hardly used in Spain and even less linked with the population registers, which is one of the novelties of our *Demodesigual* project. In 1870, three personal tax classes were created but in 1873 the tax was abolished completely for one year (Melis Maynar, 2019). It was reinstated in 1874 with nine tax classes and in 1884 11 classes were instituted that lasted for the rest of its duration. From 1874 onwards, the tax classes were based on information about direct taxes paid on real estate or industrial activities, annual salaries (only workers with annual contracts, not day laborers), and

annual housing rents (Pérez Hernández, 2020) (see Table 2). Its form was similar to that of the population registers. The register was organized on a personal basis, gathering personal, occupational, address and fiscal information for all citizens — including married women — from the age of 14 years on. All residents of a municipality were considered taxable persons.

These taxes were modest in terms of yields, but they were important in terms of the creation of a liberal tax system. They were more progressive and inclusive than previous taxes as they taxed all social strata (Marín Corbera, 2010). Preparation of the collection lists and the tax collection itself was a municipal responsibility, like the population registers, which is another indication of the quality of these sources. Its broad coverage in terms of measures of wealth and income for much of the population is also worth noting. However, active and inactive women alike fall into a single fiscal category, making its use as an economic measure unadvisable, whereas the men are represented in all the tax categories.

In mid-December 2020, the BALL contained 263,786 individual observations from the population registers and censuses from 13 different municipalities and 38,103 from the personal taxes' material from eight of them. Their population size of these cities and industrial and rural towns went from 500 to 12,000 inhabitants due to cities as Sant Feliu which showed an important urbanization as a result of industrialization and their role as head of the county (see Table Appendix 1). Other towns, as Santa Coloma de Cervelló, included an important industrial colony, where Molins de Rei and Martorell presented a relevant textile industry. The rest of the towns were mainly rural with an economy based on a commercial agriculture as a result, among other reasons, of its proximity to the city of Barcelona. The average size of households ranges from almost five to three as a clear consequence of the demographic transition, being Catalonia one of the forerunners in Spain mainly because of an early fertility decline. Actually, the whole territory covered by the BALL database — Baix Llobregat county — is part of the territory of the Diocese of Barcelona included in the BHMD, which makes both databases compatible, although they only coincide in time in the 19th century and the first years of the 20th century. At the same time, the BALL database will continue growing by adding new counties. In fact, the Maresme county is already under construction and in September 2021 began the collection of the Selva county. All these counties share a border.

Table 2 *Compiled information in the population registers, national censuses and personal taxes, 19th–20th centuries*

	Population registers/Censuses													
	1820–1829	1830–1849	1850–1859	1860–1869	1870–1879	1880–1889	1890–1899	1900–1909	1910–1919	1920–1929	1930–1939	1940–1949	1950–1965	
First name	X	X	X	X	X	X	X	X	X	X	X	X	X	
Surname 1	X	X	X	X	X	X	X	X	X	X	X	X	X	
Surname 2			X	X	X	X	X	X	X	X	X	X	X	
Occupation			X			X	X	X	X	X	X	X	X	
Sex										X	X	X	X	
Number of men in household	X													
Number of women in household	X													
Relation to the head of household						X	X	X	X	X	X	X	X	
Marital status		X	X	X		X	X	X	X	X	X	X	X	
Age		X	X	X	X	X	X	X	X	X	X	X	X	
Street name		X			X	X	X	X	X	X	X	X	X	
Citizenship										X	X	X	X	
Municipality of birth						X	X	X	X	X	X	X	X	
Province of birth						X	X	X	X	X	X	X	X	
Date of birth						X	X	X	X	X	X	X	X	
Municipality of usual residence						X	X	X	X	X	X	X	X	
Province of usual residence										X	X	X	X	
Years living in the municipality						X	X	X	X	X	X	X	X	
Months living in the municipality						X	X	X	X	X	X	X	X	
Property tax						X	X	X	X					
Industrial tax						X	X	X	X					
Literacy						X	X	X	X	X	X	X	X	
Wages and salary										X		X	X	

Table 2 *Continued*

	Personal taxes		
	1868–1879	1880–1889	1890–1943
First name	X	X	X
Surname 1	X	X	X
Surname 2	X	X	X
Age		X	X
Marital status		X	X
Occupation		X	X
Geographical origin (municipality)		X	X
Geographical origin (province)		X	X
Street	X	X	X
House number		X	X
Floor			X
Tax class	X	X	X
Number of family members	X		
Salary (annual contract)		X	X
Annual housing rent		X	X
Property and industrial tax		X	X

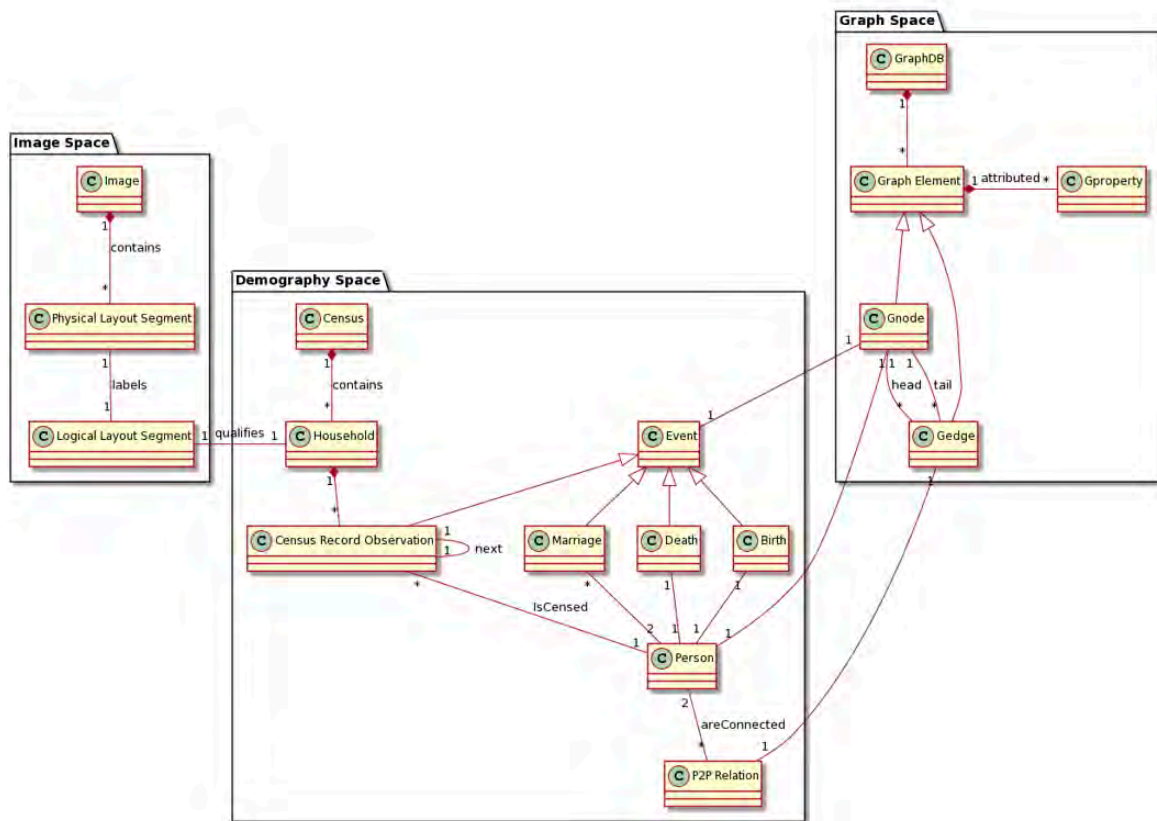
3 DATA STRUCTURE AND HARMONIZATION

3.1 DATA STRUCTURE

The information extracted from the sources (marriage license books and the population and fiscal census materials) are stored in relational databases. The demographic data of the BHMD and the BALL are complementary so the two databases share the same storage structure, allowing the information to be merged and record linkage operations to be performed to organize the data more succinctly. Record linkage operations result in discovered relations between information items. This suggests a second representational layer based on linked data (graph-like) representation. Thus, there are two levels of information using two kinds of database management systems (DBMS), namely a relational database and a graph database (see Figure 1). The relational database is the core DBMS, storing the information extracted directly from the digitized documents using the different data entry procedures: automatic information extraction using the computer vision techniques described in section 4, and the crowdsourced data collection through web-based platforms and game-sourcing mobile applications. The second level extends the original database with semantic information, implemented as a knowledge graph. A knowledge graph is a typical representation in artificial intelligence of a collection of interlinked entities. In our case, the entities are individuals, events, or places. This representation puts the data into context via semantic linking, providing a framework for data analytics and reasoning (e.g., finding communities, genealogies, life courses). Although the relational database satisfies the functional requirements, the use of a graph database, with native linked-data analytics operations, allows more efficiency and flexibility.

The complete information architecture is graphically described in Figure 1 using an object-oriented notation (a class diagram). It is made up of three abstraction layers: the image space, the demography space and the graph space. First, the image space stores the information directly related to the digitized documents and the physical/visual and logical structure of the relevant regions. The document images in a raster format are stored together with the bounding box coordinates corresponding to the segmented regions, along with the corresponding labels (logical segmentation). The image-related information is the input to the computer vision algorithms for information extraction.

Figure 1 Class diagram of the data model architecture of the BHMD and the BALL



All the information ingested from the original documents (user transcription/validation, automatic reading) is stored according to the schema presented in Figure 1 in the demography space. Its central class, *Person*, contains information that is permanent over time. For each *Person* there are different *Observations*, obtained from the analyzed censuses/marriage licenses. These observations are grouped into *Households* for the census. Each demographic document is represented as an *Event*, which can be specified as a *census record* or a *marriage record*. The proposed relational architecture has a flexible design to be able to store the information contained in the different census records as primary data, coping with the heterogeneity of different sources, time periods, and locations, and easily scaled to integrate other demographic sources like marriage, birth, or death records. This representation is compatible with the *Intermediate Data Structure* (IDS) format proposed by Alter and Mandemakers (2014). IDS is a simplified data schema that aims to map the diversity of data in European historical micro databases into a common format.

The graph space uses a graph-based model to represent a knowledge graph that infers semantics in terms of interlinked data. This knowledge graph is a conceptualization of the "historical social network" constructed from demographic sources. Using a graph representation, nodes correspond mainly to persons (observations) and events, according to the structure described in Figure 1. Graph edges represent relations between individuals (both directly extracted from the source documents, like genealogy relations, or semantic relations deduced by artificial intelligence reasoning tasks). The graph is dynamic because it evolves over time and is composed of different subgraphs. Each individual subgraph corresponds to a particular census, providing a static picture of the population at a specific time, or a marriage record. A knowledge graph also allows graph-based data analytics techniques for querying the database to be applied, including record linkage and community detection. Thus, life courses of individuals can be constructed by linking the corresponding record of the observation at time t with the observation at time $t+1$. This structure also allows future users to add specific events for individuals to achieve a more complete reconstruction of their life course. We provide further details of the knowledge graph techniques used in these databases in Section 6.

From a technical point of view, the main difference between a relational database and a graph database is the way relationships between entities are stored. Relational databases are good for highly structured data that can be organized in different tables. Relations between individual records must be defined at table level. Graph databases store relationships at individual record level. It seems an ironic, but relational databases are not good with relationships, especially when they are dynamically created. This is why we use a graph model

on top of a relational model. Relational databases are faster when handling large numbers of records because the structure of the data is known a priori. But graph databases are more efficient when handling linked data, and answering queries that involve a navigation through this linked data. We can roughly differentiate the levels saying that the relational level (demography space) stores the demographic information of persons, and the graph level (graph space) stores the contexts as linked data, which facilitates the subsequent analysis and browsing.

A side information that is stored is the metadata associated with the entities collected from the web-based crowdsourcing platform and game sourcing mobile application. This is basically the support metadata for these platforms (golden tasks, number of transcriptions per image, forms presented to the transcribers, etc.).

3.2 DATA HARMONIZATION

The BHMD has already been harmonized and the ongoing BALL database is in the process of being harmonized to make the data comparable through time and compatible with other data from similar or different national or international sources. This procedure consists of two steps: first, a linguistic standardization of names, surnames, occupations, and geographic locations, and second, a codification of the last two variables following international classifications, tasks shared among different researchers and technicians.

The name and surname standardization was a key step because the same individual could appear with seemingly different names and surnames because they were recorded using different spellings or abbreviations (Goiser & Christen, 2006; Herzog, Scheuren, & Winkler, 2007; Schürer, 2007). These issues needed to be addressed mostly for the entries in Catalan, which was not standardized until 1913 having important dialectal differences and at the same time influenced by Spanish, French, and Occitan, all of which favored the appearance of many written variations of the same name or surname (Peytaví Deixona, 2010; Rubio Vela, & Rodrigo Lizondo, 1997). To this end, we compile dictionaries with the different variations of each string variable name present in our databases (names, surnames, occupations, and geographical location) (Bloothoof, 1998).

The dictionary of first names contains 22,951 different entries, 10,667 of which are female and 12,284 male first names, all Catalan and Spanish. The one for the surnames gathers 141,129 entries in different languages, including Catalan, Spanish, French, Occitan, Italian, and English, among others. The number of spelling variations ranges from 1 till 96, with the surname *Vall* as the entry with the most variations (96). There are 24,399 items of Catalan and Spanish vocabulary relating to occupations, some of which are variations of the same ones. The dictionary of the geographical locations contains 47,559 items divided into levels corresponding to the geographical divisions of parishes/municipalities, regions, and countries. While this is an intensive task, the record linkage success rates using string distances are high (Villavicencio, Jordà, & Pujadas-Mora, 2015).

The information regarding occupations was codified according to the Historical International Classification of Occupations (HISCO), based on the International Labor Organization's (ILO) ISCO 68 classification (van Leeuwen, Maas, & Miles, 2002). Its application to the BHMD needed some adjustments due to most occupations coded belonging to the preindustrial period (Pujadas-Mora, Romeo-Marín, & Villar, 2014). Special attention was paid to the nobility titles and political institutions and a new subsidiary classification created to maintain their nature and diversity. There are hardly any women's occupations in the BHMD, although their conditions were frequently reported as slave, captive, freed slave, or prostitute, which the HISCO classification did not originally contemplate. Moreover, new codes had to be created for some controversial occupations such as gambler or bandit. The adaptation of the HISCO classification to the BALL database, however, was smoother thanks to the experience gained building the BHMD and the time span reference period of the database, the 19th and 20th centuries.

The HISCO classification also facilitates the application of other stratification schemes such as HISCLASS, HISCAM, and SOCPO (Lambert, Zijdeman, van Leeuwen, Maas, & Prandy, 2013; Van de Putte, & Miles, 2005; van Leeuwen, & Maas, 2011). The 12 categories gathered in HISCLASS were regrouped into eight classes in line with the socioeconomic characteristics of the Barcelona area. Groups 3, 4, and 6 (lower manager, lower professional and clerical sales, and foremen) and 7 and 8 (low-skilled and unskilled workers, and low-skilled farm worker and unskilled farm workers) were combined. Moreover, an extra level for the nobility was included in all three classifications due to its social class significance (score 100 on HISCAM, group 0 on HISCLASS, and group 6 on SOCPO), with the two databases totaling 565 different HISCO codes.

Last, the geographical locations were coded following the 1860 and 1991 municipal bases of the Spanish National Institute of Statistics and all the parishes were geo-referenced within the corresponding municipal

borders. For those locations outside Spain a specific country code was created. Altogether, 1,458 codes were assigned.

4 TRANSCRIBING THE DATA BY WAY OF CROWDSOURCING AND GAME SOURCING

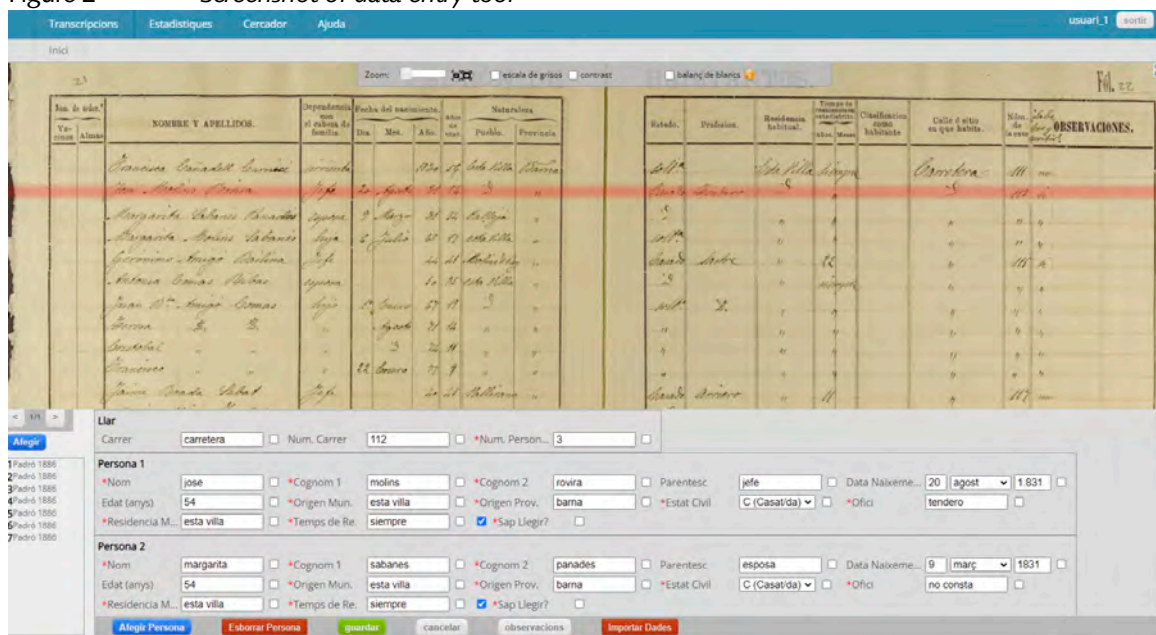
Manual transcription from original sources using a computer or not has been the system traditionally used since the mid-20th century. This way of constructing demographic databases has proven to be tedious and time-consuming. To speed up and to facilitate the data entry process of the BHMD and the BALL database, we have developed several computing alternatives based on crowdsourcing and game sourcing paradigms. Secondly, we used computer vision algorithms to extract the information contained in the marriage licenses, population registers, national censuses, and personal taxes.

4.1 MANUAL TRANSCRIPTION USING CROWDSOURCING PLATFORMS

For this first approach, we developed several web-based platforms (Fornés, Lladós, Mas, Pujadas-Mora, & Cabré, 2014) to extract the information contained in the demographic-like sources. The BHMD and the BALL database have their own platforms, which are used by many users in parallel, following the crowd-based outsourcing concept. This consists in dividing a long task into many smaller tasks and distributing them among a large group of people, particularly from the online community, thereby shortening the time required and also allowing a higher volume of data covering larger geographical areas to be gathered. The user interface of our web-based platform (Fornés et al., 2014) consists in a data entry tool with a user-friendly environment. This means that both the scan of the source and the data entry form are combined in one view (see Figure 2). The web incorporated several tools to improve legibility such as zoom and image contrast. This is particularly useful in the case of paper degradation, as has happened with one of our sources. The iron gall ink used having corroded some parts of the original documents, mainly from the 16th century. The BHMD has been compiled entirely using a web-platform. For its part, the BALL platform incorporates a computer assisted transcription system, as explained in the following subsection.

A total of 153 paid transcribers (69 men and 84 women) participated in building the BHMD from backgrounds as diverse as secondary school pupils to university students, consultants in history, local historians, retired people, and PhD holders in history. Volunteers totaling 175 (81 men and 94 women) collaborated in constructing the BALL, almost all of which were local history and/or genealogy enthusiasts with previous knowledge of the source materials they were transcribing. They were from diverse cultural backgrounds, were mainly university graduates, and had a mean age of 59 years.

Figure 2 Screenshot of data entry tool



4.2 SEMI-INTERACTIVE TRANSCRIPTION SYSTEM

Since population registers and census records are recorded in intervals of a few years, the individuals in each household are quite stable. Hence, once the records for a specific year have been transcribed, this redundancy can be exploited to assist the transcription. The main idea behind the semi-interactive transcription system we developed (Mas, Fornés, & Lladós, 2016) is to transfer the individuals in one previous population register/census to the next one (see Figure 3). The computer vision algorithm first detects and transcribes the street address in the original manuscript page. The system then accesses the database and retrieves the list of individuals that were living in that household in the previous population register/census. In case the transcribed street address is not found in the previous census, it means that no previous information is available on that particular household, so the data must be manually introduced.

In case the household is found in the previous census, the system proceeds to check which individuals still appear in the current one. The search for each individual (name and surnames) in the new population register/census is performed using a word spotting algorithm. Word spotting consists in finding a particular word in a manuscript that has not been transcribed (using OCR or similar). This means that the search is performed directly on the digitized image. Since the task consists in searching concrete words (e.g., names and surnames), the method is more accurate than automatic transcription.

So, when the word spotting finds the name and surnames of the individual, it means that this particular individual appears in both the previous and the current census (so, no change in address). So, the individual's information from the previous census is copied on the data entry form of the current census. The user simply needs to update the still living individuals with their age or marital status, complete the new fields (where appropriate), transcribe the new members of the households (birth, marriage, or immigration), and delete the individuals who emigrated or died between two consecutive rounds.

4.3 AUTOMATIC TRANSCRIPTION USING COMPUTER VISION

4.3.1 THE HANDWRITTEN TEXT SYSTEM

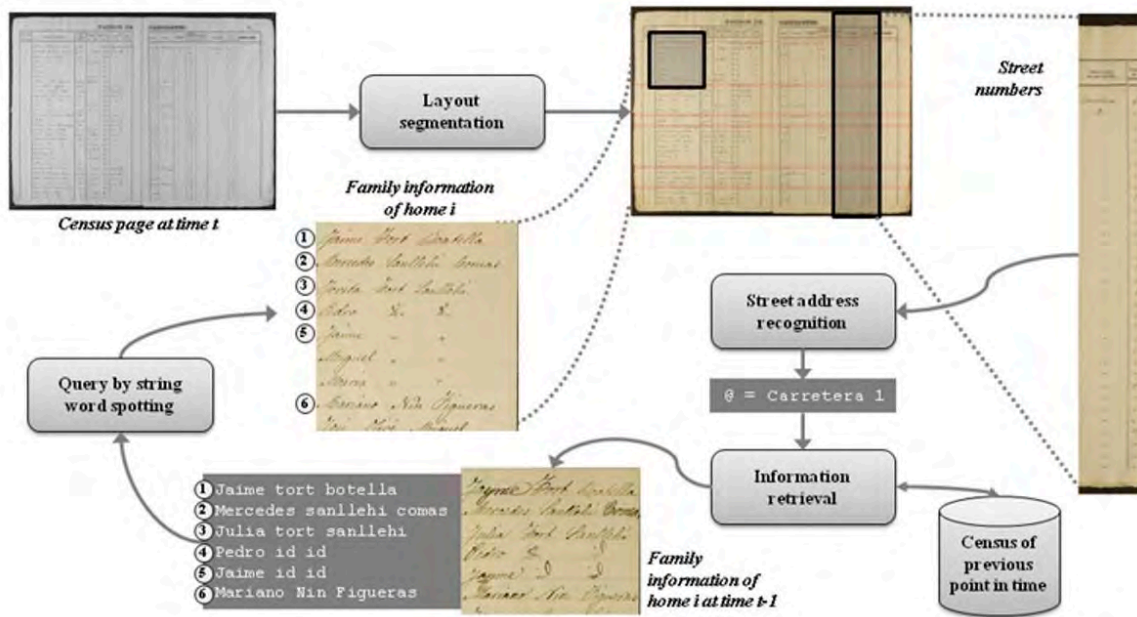
In the second approach, we assume that there is enough labelled (transcribed) data from the sources to be able to use computer vision and document image recognition techniques. We used a Handwritten Text Recognition (HTR) system (Toledo et al., 2017) trained on these kinds of documents (e.g., similar structure, same time period, similar handwriting style) and based on deep learning architectures, or more specifically on recurrent neural networks.

Since the marriage license records are written in blocks, once the different blocks and text lines have been automatically segmented and transcribed, a named entity recognition algorithm (Toledo et al., 2019) is used to extract names, surnames, occupations, places, etc.

For the population register/census documents, the steps of the automatic recognition system are:

1. A layout analysis algorithm detects the different columns and rows in the tabular census documents. Each cell in the table is then segmented into words.
2. For each column, a word image clustering method groups words by visual similarity. Thus, the groupings will likely contain instances of the same name, surname, etc.
3. The most populated clusters are selected, which correspond to the most common names, surnames, etc.
4. The words that belong to each one of the populated clusters and transcribed using a Handwritten Text Recognizer algorithm.
5. Since the system knows the exact location in the page (this means the column and row in the table) for each word in the clusters, the transcription of each word is introduced in the corresponding field in the form (e.g. if a cluster contains many instances of a common surname, all individuals with this specific surname will be filled in the form, in one single step).

Figure 3 Examples of the semi-interactive transcription system



Source: Mas, Fornés and Lladós (2016).

Figure 4 Example of text confirmation by way of videogaming



Given that the accuracy and performance of all these methods depends highly on the amount of labelled data to train the models, differences in handwriting styles, and degradation of the paper, the transcriptions must be validated manually. Instead of using a crowdsourcing web-based platform, we opted to incorporate gamification principles to engage users in the validation, commonly referred as game sourcing. Game sourcing consists in the application of game design elements into non-game contexts, like crowdsourcing. More specifically, we have developed several video games for mobile devices (Chen et al., 2018). These mini games (see Figure 4), grouped together in a video game named *WordHunter*, are designed to engage users

in validating the transcriptions and confirming the computer vision algorithm output applied to population registers/censuses or personal taxes. The three games are:

- a) *WordHunter-Difference*: A game to validate the word clustering algorithm, consisting in word images contained in a given cluster, which the user must identify as being the same or not.
- b) *WordHunter-Match*: A game to validate the transcription algorithm. A set of word images are shown on the screen along with the most probable automatic transcriptions provided by the HTR system, from which the user must select the correct answer.
- c) *WordHunter-Jump*: A third game, also used to validate the transcriptions, which emulates a platformer game. The user makes the videogame character jump to catch the correctly transcribed words by the HTR system.

More than 40 volunteers participated in the validation using videogames. Of these, 28 were men and 14 were women, 32 were experts (with a background in history and/or paleography) and 8 had no previous knowledge of the material. The videogames have been available for download from Google Play for the last 6 months and almost 100 users have already played them.

4.3.2 STEPS OF THE AUTOMATIC TEXT RECOGNITION SYSTEM

Layout Analysis

Layout analysis is crucial for the semantic interpretation of a document. However, layout usually depends on signals in the image (e.g., lines that define the rows, columns, and cells of a table), which serve to define the limits between the different parts of the document. In some cases, such as the marriage records of the BHMD, the entries are recorded in blocks rather than tables. In others, the signals are not sufficiently clear due to the quality of the image or degradation of the paper. For these reasons, the layout analysis module must use this incomplete information to understand the logical structure and extract the page layout. Consequently, it is important to study how the maximum layout information can be obtained from the parts of the document where there are clear image signals, using automatic inference to determine the coherent parts of the document according to these clues (Kleber, Diem, Dejean, Meunier, & Lang, 2018). In addition, the recent success of graph neural networks for semantic segmentation in scene images makes them promising architectures for the analysis of complex population documents (Carbonell, Riba, Villegas, Fornés, & Lladós, 2021).

Handwritten Text Recognition

Once the layout of the document has been recognized, the next step in the extraction of information from digitized documents is the recognition of text. Although Optical Character Recognition is quite accurate for recognizing printed text, the recognition of handwritten text is still a challenge. Even though Handwritten Text Recognition (HTR) systems based on deep learning models perform better than traditional approaches based on Hidden Markov Models or Recurrent Neural Networks, the accuracy is still unsatisfactory regarding historical manuscripts. The difficulties in processing historical handwritten documents are mainly due to paper degradation, different handwriting styles, old vocabulary, etc., causing frequent errors in the transcription.

Research on more robust and generic HTR models is therefore required. Some recent deep learning approaches based on sequence-to-sequence models (Kang, Riba, Villegas, Fornés, & Rusiñol, 2020a) and transformer networks (Kang, Riba, Rusiñol, Fornés, & Villegas, 2020b) have been proposed. These models have significantly improved on the HTR performance of previous architectures, but the large differences in the handwriting styles in documents from different regions or centuries still pose a problem. Therefore, there is a need for research into models that can adapt to unseen handwriting styles, meaning that techniques for domain adaptation and transfer learning should be explored. Some attempts have already been made towards unsupervised writer style adaptation (Kang, Rusiñol, Fornés, Riba, & Villegas, 2020c) to recognize segmented handwritten words, but these techniques need to be further explored and extended to deal with full sentences.

Named Entity Recognition

Analysis of the document layout and the transcription of text is not enough to fill the corresponding knowledge database, the reason being that the transcribed text needs to be semantically labelled, providing a semantic meaning. In other words, the named entities (e.g., names, surnames, locations, dates) must be

detected so that the information contained in the documents can be properly entered in the way required by the database.

The typical procedure consists in applying Named Entity Recognition models from the Natural Language Processing field to the transcribed text. However, the transcribed text contains errors, which will affect the performance of the subsequent Named Entity Recognition module. Consequently, there have been some attempts in recent years to perform text detection, text recognition, and named entity categorization in an integrated neural network model (Carbonell, Fornés, Villegas, & Lladós, 2020), commonly referred as multi-task learning. This joint architecture must therefore learn to perform several tasks at the same time, which can potentially improve the overall performance and reduce errors. However, this kind of approach consists in large deep learning architectures that need many labelled documents to train. Since labelled data is costly to produce and barely available, the generation of "realistic" synthetic documents (Das et al., 2020) to increase the amount of training data is worth exploring. All in all, the automatic extraction of information from images of population documents have shown to speed up the data entry process, although the performance of such techniques is not perfect, so a manual validation is still needed. In general, the performance of our developed methods vary depending on the paper degradation and the handwriting style. For example, the transcription accuracy varies from 60–90%, whereas the word spotting method reaches a performance between 80–90%. These values demonstrate that, although beneficial, automatic tools need human supervision for quality check. For this reason, further research will be focused on improving the automatic analysis of the document layout, the recognition of handwritten text, and the recognition of named entities.

5 DATA LINKAGE

Record linkage as the linkage of different appearances of the same persons is necessary to reconstruct the life course of individuals. This happened with both databases, in the BALL database to reconstruct families from a multigenerational perspective and to reconstruct a succession of marriages forming lineages, as is the case with the BHMD. The patrilineality of the transmission of surnames, which prevailed from the medieval period onwards in Catalonia mostly for men, and not changing female surnames throughout their life from modern times onwards, facilitated enormously the reconstruction of families and the application of different approaches for the data linkage. The varying number of reconstructed generations can be explained not only by the social and biological success of their descendants but also by migration. However, loss of migration is not that problematic, given the size of the territory and the continuity of the sources. At the same time, the areas covered by the two databases have historically been poles of attraction. We describe the three different types of methods we developed for linking the data and creating the social network, including constructing pedigrees and life courses. These three methods cover from the most basic statistical methods, which are based on string distances, to more sophisticated methods which build intermediate representations that allow record linkage between heterogeneous sources of information. The first proposed method, based on string distances, and the second one based on visual word search (namely word spotting) have been applied to both databases. The third method, which is based on knowledge graphs, has been applied only to the BALL database.

5.1 STRING DISTANCES

The first method is based on the comparison of strings (e.g., names and surnames). In the case of the marriage license records contained in the BHMD, the objective was to recursively link the married couples with their corresponding parents following a backward approach, thus creating lineages as pedigrees. For this purpose, we developed a software application that runs three different similarity measures between strings. These algorithms consider the length of each string and the position of each letter. The string comparison is based on the "Levenshtein distance" and a combination of the so-called "bag distance" and the "longest common substring distances". These algorithms are adapted to Catalan phonetics, allowing spelling variations. This meant that we could define which letters of the alphabet were usually used in the same place (similar phonetics) so that the similarity algorithm could take them into account when computing the final string distance. Besides the string distance, we also imposed several restrictions for linking the data for the BHMD when lineages were created. These restrictions consider data plausibility between marriages, setting a minimum difference of 15 years between the marriages of parents and children, and a 50-year maximum difference in extreme cases. To minimize the overlinks due to the similarity of onomastics, we carried out

a complete cross check of the links following historical and probabilistic criteria and considering both the number of kilometers between the geographical location and the social distance between the marriages of parents and children. The spatial distance between the place where the marriages of parents and children took place should follow a logical sequence according to the mobility patterns of the study period. The occupations of parents and their descendants and the amount of tax paid following each marriage should not be more than two tiers different in the fee scale described.

By applying the software and the aforementioned rules, family trees have been created for two different periods, since until 1481 the filiation of the bride and for the period 1645–1753 the name of the parents were not recorded, being key for the linkage of the generations. The first period ranges from 1500 to 1643 and the second period, from 1715 to 1880. Among the links generated by our rules and software (with a threshold of 85%), 33% of the total were validated. This reflects the fact that the Barcelona area was (and still is) a dynamic and mobile place and the restrictive nature of the method. The maximum number of linked generations for the 16th–17th centuries is four and for the 18th–19th centuries it is six, which could be explained both by a higher survival and greater stability of the surnaming system. Nevertheless, in both periods, two-level generations (parents–children) are the most frequent (see Table 3).

In the case of the BALL database, individuals were linked according to the string edit distance, also with a threshold of 85%. The procedure and restrictions differed slightly from those mentioned above for the BHMD, with the individuals in each household linked according to the relationship with the head of the household (e.g., sisters/brothers, sons). When a new census is introduced in the database, the algorithm searches whether these individuals also appear in the previous census. The string edit distance is computed together with some restrictions, namely that the street address should be the same and the individual's age should increase in a coherent way. If these constraints are satisfied, the unification is fully automatic and the new relatives are linked (e.g., newborns). By December 2020, all individuals present in the national censuses and censuses of five of the 14 municipalities (Sant Feliu de Llobregat, Collbató, Castellví de Rosanes, Santa Coloma de Cervelló and Sant Vicenç dels Horts) comprising the BALL database have been linked both across time and space. Thus, on average, more than 80% of the individuals have been linked to create a unique person. This implies that 80% of the individuals present in these municipalities have at least two observations over time either in the same municipality or in different municipalities (see Table 4). The maximum number of observations for the same individual is 12. In addition, tax information from personal taxes has also been linked, with a total linkage ratio of 95%.

In the coming months, new municipalities will continue to be linked to the BALL database. But it is also planned to interconnect the BHMD and BALL database for the overlapping years — 19th century and early 20th century — as the territory is totally coincident since the municipalities of the Baix Llobregat county were part of the Diocese of Barcelona, ecclesiastical demarcation of the BHMD, as explained above.

Table 3 *Number of marriage licenses, linkage validation and number of generations linked in the Barcelona Historical Marriage Database*

BHMD	XVI–XVII (1500–1643)	XVIII–XIX (1715–1880)
Marriage licenses	125,140	334,011
Linked marriages	21,200	109,224
Genealogies		
2 generations	19,425	95
3 generations	172	12
4 generations	55	2
5 generations		255
6 generations		11

Table 4 *Number of inhabitants and linked individuals between two consecutive population register/census in BALL database*

Sant Feliu de Llobregat											
	1878	1881	1889	1906	1910	1915	1920	1924	1930	1936	1940
Total inhabitants	2,747	3,002	3,118	3,606	3,807	4,329	4,352	5,569	6,383	7,020	6,720
Total households	673	641	645	804	867	937	923	1,048	1,162	1,458	1,678
Linked individuals	-	2,573	2,627	3,169	3,411	3,788	3,888	4,544	5,295	5,940	5,109
% Linked individuals	-	85.7%	84.3%	87.9%	89.6%	87.5%	89.3%	81.6%	83.0%	84.6%	76.0%
Persons by household	4.08	4.68	4.83	4.49	4.39	4.62	4.72	5.31	5.49	4.81	4.00

Santa Coloma de Cervelló						
	1901	1924	1936	1940	1945	1950
Total inhabitants	542	1,132	1,311	1,218	1,167	1,222
Total households	127	260	325	329	307	308
Linked individuals	-	805	1,122	1,124	1,063	984
% Linked individuals	-	71.1%	85.6%	92.3%	91.1%	80.5%
Persons by household	4.27	4.35	4.03	3.70	3.80	3.97

Sant Vicenç dels Horts								
	1921	1925	1930	1935	1940	1946	1950	1955
Total inhabitants	2,097	1,985	2,950	3,129	2,980	3,014	3,323	3,717
Total households	535	409	783	716	763	791	1,030	1,168
Linked individuals	-	1,844	2,560	2,809	2,721	2,713	2,911	2,676
% Linked individuals	-	92.9%	86.8%	89.8%	91.3%	90.0%	87.6%	72.0%
Persons by household	3.92	4.85	3.77	4.37	3.91	3.81	3.23	3.18

Callbató														
	1887	1889	1896	1900	1905	1910	1916	1920	1924	1930	1936	1940	1945	1950
Total inhabitants	781	812	686	703	665	809	796	560	754	531	525	474	425	418
Total households	172	170	147	182	144	162	166	141	170	138	138	162	125	147
Linked individuals	-	651	647	607	634	716	748	544	650	516	503	440	399	352
% Linked individuals	-	80.2%	94.3%	86.3%	95.3%	88.5%	94.0%	97.1%	86.2%	97.2%	95.8%	92.8%	93.9%	84.2%
Persons by household	4.54	4.78	4.67	3.86	4.62	4.99	4.80	3.97	4.44	3.85	3.80	2.93	3.40	2.84

Castellví de Rosanes						
	1924	1930	1936	1940	1945	1950
Total inhabitants	283	279	273	281	242	268
Total households	61	64	62	73	50	60
Linked individuals	-	260	252	236	222	195
% Linked individuals	-	93.2%	92.3%	84.0%	91.7%	72.8%
Persons by household	4.64	4.36	4.40	3.85	4.84	4.47

5.2 VISUAL WORD SEARCH

The visual word search corresponds to the word spotting procedure described in the semi-interactive transcription system section. As explained previously, the key is to benefit from the redundant information between consecutive census records, which can be used to speed up the transcription while at the same time unifying the individuals that appear in the population register/census, and involves linking the same individual across different censuses, thus creating the life course. For each household in a previous census, the word spotting algorithm searches each individual living in that household in the current census. This search is performed by visual similarity, which means that the algorithm internally creates a visual representation of each name and surname. The individual is considered found if the system finds the same combination of name and surname/s, and the age is feasible. In such cases, the linkage (i.e., unification) is automatically performed. This procedure has proven to reduce the time of transcription of a new population register or census by 70% (Mas et al., 2016).

5.3 KNOWLEDGE GRAPH-BASED METHODS

A knowledge graph (KG) is a data representation that uses graphs to emphasize the links between data elements (Schneider, 1973). A KG formally represents object semantics by describing their relationships. Moreover, they can also make use of ontologies in an abstraction layer to allow logical inference for implicit knowledge retrieval (Baader, Calvanese, McGuinness, Nardi, & Patel-Schneider, 2003). The KG has become more popular since Google introduced their KG to complement their string-based search engine (Singhal, 2012), and they are currently used by many multinational firms to provide their customers with tailored services. There are two major KG representations: entity-relation (ER) graphs and entity-attribute-relation (EAR) graphs (see Figure 5). In an ER graph, nodes are the world objects (entities), which contain a set of attributes that describe each node, and edges (relations), which link nodes between them. This is the simplest KG representation and any AI algorithm reasoning on it must navigate through the graph edges to infer a result. However, databases like the BHMD and the BALL database, contain slightly different information from individuals that cannot be modelled by a single ER KG. EAR graphs model this data better than ER graphs because there are two kinds of nodes: entities and attributes. The node attributes in an ER KG are extracted from the node and are "promoted" to nodes in the EAR KG and, consequently, each attribute node is connected to its original node in the equivalent ER KG representation. In Figure 5 we show how this procedure works. For instance, individual information like "first_name", "surname", "marital_status", etc. that are defined as attributes of individual nodes in Figure 5a are defined as nodes in the EAR KG representation which are linked to individuals in Figure 5b. Thus, an individual is not represented only by a set of attributes but by a subgraph, in which the central node is the Individual. This representation provides more flexibility as kin relations and individual information are represented homogeneously and all of them contribute to build the numerical representation that will be used to represent individuals for record linkage tasks.

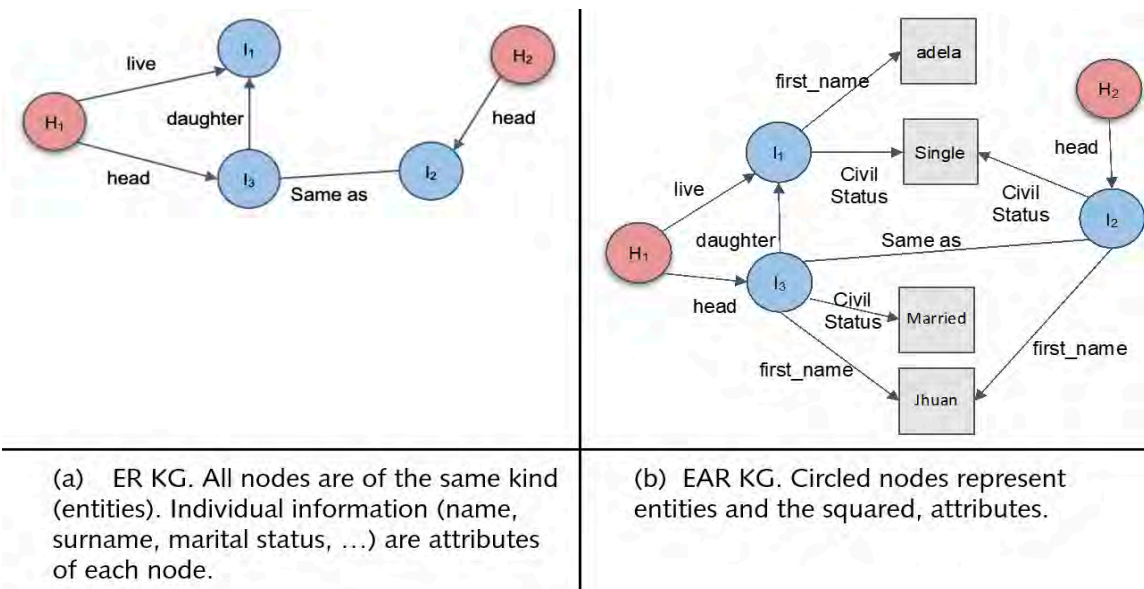
As motivated above, we use EAR KG to model data extracted from the BALL database. Following this representation, entities mainly correspond to individuals and households, while attribute nodes are the field values (name, surname, age, occupation, household address, etc.) describing each individual or household. The KG also has two kinds of edges. On the one hand, there are the edges linking entity nodes representing family relationships between individuals living in the same household, the edges linking individuals and households to denote the household in which each individual lived, and the edges between two or more individual nodes representing the same person through their records over years of censuses. On the other hand, there are also edges between entity nodes and attribute nodes. As mentioned above, attribute nodes contain the values of the fields extracted from the BALL dataset. All the attribute nodes are therefore essentially composed of the word vocabulary appearing in this database. The edges between entities and attributes represent the semantics, i.e., name, surname, occupation, and so on, assigning the item to each attribute node when it is linked to an entity.

Record linkage on an EAR KG is an inference method to discover links between individual entities representing the same person over time. Latest advances in this field compute numerical representations of each node in the KG. Each numerical representation is a set of N real values, where N is an arbitrary value set beforehand by an expert and usually takes a value ranging from 32 to 64, that encodes each node. Since each node in this database represents, either an individual or a household, each numerical representation is built from the attribute nodes that are linked to each individual or household. The main differences between these methods are the way that these numerical representations are computed for each individual, but all of them essentially seek the same purpose, for the two nodes that represent the same individual, or household, to

have their numerical representation close to each other. For instance in Figure 5b, we can see two individuals, I2 and I3, that should be identified as being the same person by any record linkage method. The proposed EAR KG methods will learn numerical representations that will be encoded in a N-dimensional space. These representations will consider the edges that connect nodes I2 and I3, respectively. If both nodes had the same neighbors, i.e. they live in the same household, had the same first name, marital status, etc. they would exactly have the same numerical representations and hence the record linkage method would provide a perfect match. In the given example, I2 lives in a different household and his marital status has also change from "single" to "married". Consequently, the numerical representations will differ but should be close enough to provide a match of the record linkage method.

We have applied a representative set of these techniques to a subset of the BALL, choosing records from the municipality of Sant Feliu de Llobregat and splitting the data into the following partitions. We have imported the data into a local Neo4j Database server, which is a native graph database system. We used two pairs of datasets for training, (A,B) as (1889, 1906) and (1930, 1936), (1906, 1910) and (1936, 1940) for validation, and (1910, 1924) and (1924, 1930) for testing. The selected methods are TransE (Bordes, Usunier, Garcia-Duran, Weston, & Yakhnenko, 2013) and TransH (Wang, Zhang, Feng, & Chen, 2014), which are applied to ER KG; SEEA (Guan et al., 2017) and KR-EAR (Lin, Liu, & Sun, 2016), which are applied to EAR KG; and the WERL method, which is applied to an extension of the EAR KG that we recently developed in (Gautam, Ramos-Terrades, Pujadas-Mora, & Valls-Fígols, 2020).

Figure 5 Types of KG representations



Source: Gautam, Ramos-Terrades, Pujadas-Mora, and Valls (2020).

Table 5 Record linkage results for the BALL database only for the town of Sant Feliu de Llobregat

Method	Accuracy	Precision	Recall	F-Score
TransE (ER)	0.87	0.19	0.25	0.21
TransH (ER)	0.81	0.12	0.26	0.17
KR-EAR (EAR)	0.79	0.13	0.30	0.18
SEEA (EAR)	0.91	0.07	0.10	0.08
WERL (ER)	0.99	0.99	0.67	0.80
WERL (EKG)	0.99	0.98	0.89	0.93

Source: Gautam, Ramos-Terrades, Pujadas-Mora, and Valls (2020).

Comparing benchmark methods, the reported results show a significant performance improvement on the WERL-based method in terms of the F-Score (see Table 5). The F-Score is a performance measure, ranging between 0 to 1, that combines the precision and recall of a system like a record linkage method. It is formally defined as the harmonic mean of the Precision and the Recall measures and is easily computed as:

$$F - Score = 2 \frac{Precision \times Recall}{Precision + Recall}$$

The Precision and Recall, aka sensitivity, are two measures used to evaluate the impact of false positives (Type I error) and false negatives (Type II error) in the system performance:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

where TP, FP and FN stand for true positive pairs, false positive pairs and false negative pairs, respectively. This improvement is weakening when we observe the accuracy of the methods. A few comments are needed to better understand these figures. First, record linkage is a highly unbalanced classification problem, meaning that there are many more pairs of not linked records (true negative pairs) than records to be linked (true positive pairs). Thus, automatic record linkage methods biased to not link candidates of pairs are likely to increase their accuracy. They therefore achieve low precision and recall values and, consequently, low F-score values. Conversely, the proposed WERL methods have been shown to have good accuracy and precision scores, meaning that linked pairs are likely to be correct by relatively low recall scores, and indicating that some record links are missed. Given the amount of data to be processed, the challenge is to increase the recall score while losing minimal precision.

6 CONTRIBUTIONS TO HISTORICAL DEMOGRAPHY

The historical lines of research on demography using the BHMD and the BALL deal with the intergenerational transmission of social status, intermarriage and kinship marriages within the social reproduction process (15th–20th centuries), the evolution of migratory flows in a long-term view (15th–20th centuries), and the estimation of economic inequality within preindustrial and industrial periods.

6.1 SOCIAL REPRODUCTION

Social reproduction is a classic research topic in historical demography, referring to the economic, institutional, legal, political and/or cultural mechanisms used by individuals, families, or social groups to maintain, improve and/or transmit their social position (Bourdieu, 2013). However, dimensions such as the intergenerational transmission of social outcomes (mostly of occupations or social status) and intermarriage or kin marriages, which appear to be fundamental to social dynamics and change, have been scarcely studied for preindustrial periods or from a long-term perspective. This is, in part, owing to the non-availability of sources and data for earlier periods and different historiographical traditions in every country.

Families used intergenerational transmission, intermarriage, or kin marriages, which were subject to demography, legal systems, and economy expansion and contraction (Berkner & Mendels, 1976; Ganzeboom, de Graaf, & Treiman, 1991), to place their members in better-off positions and to perpetuate their lineages biologically and socially over time. This phenomenon led preindustrial society to appear static, although it is true that levels of social transmission between generations or social endogamy, and even territorial endogamy, remained high in these societies. However, this may have varied among different social/migratory groups since they might have adapted their strategies depending on their specific circumstances and the general context (Bourdieu, 1976; Laslett, 1983), which ultimately involves group identities and consequently class formation (Kocka, 1984).

The socio-occupational destinations of sons and daughters of peasants and artisans in the area of Barcelona in the period 1547 to 1643 were analyzed using the BHMD to explore how intergenerational transmission worked in preindustrial periods (Pujadas-Mora, Brea-Martínez, Jordà, & Cabré, 2018). Destinations are understood as a combination of both ascription and achievement, i.e., the influence of parents' socioeconomic

positions and individual improvements in relation to family background. Notably, the legal context, and specifically inheritance systems, can strengthen the intergenerational transmission of occupations or status, particularly when the system is based on the principle of impartibility, which usually granted eldest sons the privilege of being heir, as was the case in Catalonia. At the same time, this research aims to shed light on the competitive or cooperative strategies adopted by families which could prevent or facilitate parent-offspring conflict (Schlomer, Ellis, & Garber, 2010; Trivers, 1974) and resource dilution (Low & Clarke, 1991; Shenk et al., 2010). A multilevel approach was applied using hierarchical linear models and logistic regressions. The benefit of this kind of approach lies in the possibility of analyzing how similar siblings were in status attainment in view of the relatively high level of ascription in pre-industrial periods.

According to the findings, high levels of intergenerational status inheritance and ascription were closely linked to the effect of the parents' social class, a plausible fact in an ordered society with an inheritance system based on the impartibility of property. The pattern among farmers, whose children also tended to be farmers, accounts in part for the high levels of intergenerational transmission. Artisans also contributed to raising the total intergenerational transmission, given that sons of artisans mainly inherited the condition but not the occupation. First-married children were the main inheritors of parental status in all social groups, and especially among farmers and artisans. However, some second- or third-married farmers' children were able to follow artisan careers. Likewise, the farmers (mainly non-inheritors) joined the then-existing waged rural labor market, which could sometimes be used to complement earnings from their own land. However, low availability of cropland (Alvarez Nogal & Prados de la Escosura, 2007; Ferrer-Alós, 1983, 2014; Vilar, 1986), accompanied by the expansion of the textile sector (García Espuche, 1998; Torras, 1998) in early modern Catalonia could have prompted families to encourage some of their children (usually the non-first-married) to take up artisan activities. Hence, rather than a model of sibling competition for family resources, we see a pattern of family economic diversification, which is one of the most relevant and novel conclusions. The most paradigmatic example of this is probably the significant number of farmers with children who became *pelaires*, or fiber preparers.

Using the BHMD, intermarriage is approached through the study of immigrant marriage due to the important migratory flow to Catalonia from France that took place mainly in modern centuries (Amengual-Bibiloni & Pujadas-Mora, 2020). The analysis of this flow has an important tradition in Spanish historiography, starting with the seminal works of Moreu Rey (1959) and Nadal and Giralt (1960), although it is little known in the international literature. The volume and length of this flow makes it one of the most important in modern Europe, one result of which is the important number of these migrants who got married in Catalonia, a neglected topic in both national and international historiography. Up to 25% of the marriages that took place in the Diocese of Barcelona in the period 1568–1639 were between a French groom and a native wife. Multivariate logistic models confirm that French grooms were significantly less likely to get married to a bride from an equivalent social group than natives, with high odds of marrying a widow bride. Nevertheless, the interaction between a specific period (the whole period, 1568–1639, was split into sub-periods of ten years) and geographical origin of the groom (native and French) presents a decrease over time of the odds of finding a partner of the same social status for French and native grooms. On the one hand, the increase in the relative weight of French grooms modified the structure of the whole marriage market, and on the other hand, native brides, mainly widows, became less reluctant to get married with a French groom. Notably, group size is a key factor to better understand the social homogamy: the larger the size, the greater the probability of homogamy (except for the nobility), even when the marriage market is considered as a segmented market by social group, geographical origin, and period. The important presence of French migrants on the marriage market shows that there was an important marriage squeeze due to more male candidates, which would explain why native widows were remarrying more than in previous or subsequent periods.

Last, the practice of kin marriages as a mode of social reproduction was explored through the BHMD. The high increase in the number of consanguineous and affine marriages throughout the 19th century was a common phenomenon in different European countries (Bittles & Black, 2010; Bras, Kok, & Mandemakers, 2010; Chacón Jiménez & Hernández Franco, 1992; Delille, 1994; Gouesse, 1984; Sutter, 1968). Two types of kinship-affine marriages, simultaneous marriages of two or more couples of siblings and marriages of widowers or widows with the sister or brother of their deceased spouses, phenomena known as levirate and sororate marriages, are considered (Pujadas-Mora, Brea-Martínez, Jordà, & Cabré, 2018). The aim of this study was to establish the social profile of these types of marriages and to try to find the determinants of choosing a partner among kin in the Barcelona area in the period 1780–1880. The farmers and skilled and lower skilled workers, mainly artisans, and the day laborers were the groups with the highest percentage of simultaneous marriages, and particularly notable was the weight of farmers in simultaneous marriages of brothers and sisters. This could be because the eldest son, the universal inheritor, was marrying at the same

time as one of his sisters, who was marrying another universal inheritor, meaning that she was marrying upwards. The high proportion of simultaneous marriages among day laborers may be due to the broad application of that occupational title. Men were classified as day laborers even when their father or brother was a land owner.

Logistic regressions were performed to disentangle the weight of the factors that determined sororate and levirate marriages, with the novelty of a control variable for the frequency (distribution) of surnames in the population. The widowers from the highest social groups, mostly elites and professionals, showed higher odds of marrying their sisters-in-law than the day laborers. Furthermore, there was no difference in this practice between urban and rural areas. An important concern in this study is the fact that these kinds of marriages were completely dependent on the demographic circumstances of the death of a spouse, the existence of an available sister or brother for sororate or levirate marriage, or the survival of at least two siblings of a marriageable age. The demographic transition helped the occurrence of these kinds of marriages due to an increased number of relatives on the marriage market following an increase in survival at marriageable ages.

6.2 MIGRATION

The analysis of historical migrations is a difficult issue due to the scarcity or even lack of specific sources that compile information about migratory events. This is the reason why onomastics, and mainly the exploration of family names, has been used in demography, history, and biology to assess the mobility of historical populations. In medieval and modern Europe, onomastics in general and particularly surnames were subject to various changes resulting from the continuous flow of the rural population to the urban centers, in addition to the homeless, pilgrims, and traders, and in modern Catalonia to a lesser extent, immigrants from the Kingdom of France. The combination of all these processes meant that between medieval and modern times there were important variations in the number, prevalence, and distribution of Catalan surnames. A total of 47% of the surnames appear for the first time in the Barcelona Diocese's Marriage Licenses Books, compiled in the BHMD, between the periods 1451–1487 and 1487–1500, which could be a sign that these surnames were brought to the area of Barcelona by recent immigrants. Conversely, 64% of surnames disappeared in the same period. These results suggest that approximately half of the surnames recorded in the BHMD during these years could have a migrant origin (Jordà, Amejérias-Alonso, & Pujadas-Mora, 2018; Jordà, Pujadas-Mora, & Cabré, 2016).

When an isonomic analysis of onomastics is carried out for the whole of Catalonia using the household counts of 1497 and 1553, three different internal migratory patterns are observed. The first is a north-south population flow, seemingly because of a steady migration from the Catalan Pyrenees (north) to the Tarragona and Montblanc areas (south). The second is an inland migration trend. And the third is a continuous settlement pattern in the Barcelona hinterland. Moreover, it should be noted that a significant increase in the number of households is observed from 1497 to 1553, which would have led to a dramatic reduction in the distribution of ancient Catalan family names (Jordà et al., 2016). These results could support the notion that during the first half of the 16th century the territory underwent several internal migratory flows, favored by the economic growth of the time and the resulting increase in the demand for labor, which was also in part covered by migrants from France (Amengual-Bibiloni & Pujadas-Mora, 2016).

Moreover, it should be noted that Catalan onomastics are the result of the interrelations between Catalan, Occitan, French, and Castilian surnames. Immigration from France and then from the rest of the different Spanish kingdoms enriched Catalan onomastics by introducing new forms of surnames, not always distinguishable from a linguistic criterion given a relatively high number of homonymic surnames among the languages and the absence of standardized languages. Cluster analysis allowed for organizing the surnames into 3 different groups by similarities of growth, independent of linguistic and historical origins. The first group shows the highest number of Catalan surnames; the second has the highest percentage of surnames borne by foreigners; and the third, which is the most heterogeneous of all the groups, comprises the largest number of Castilian surnames and the smallest number, in relative terms, of surnames influenced by French immigration (Jordà et al., 2016).

All these studies contribute to confirming that there was much geographical mobility throughout the preindustrial period, both internally and internationally, allowing a better understanding of how migrants contributed to revitalizing the cities from a demographic point of view, which had suffered a chronic negative natural growth rate due to poor living conditions, as had the Catalan coastal villages. At the same time, they help to demystify the ethnic origin of surnames.

6.3 ECONOMIC INEQUALITY

Increased economic inequality is one of the most worrying problems nowadays and has evolved into a central topic of historiographical debate. A historical perspective shows the roots of current inequality and is key in the understanding of social mobility trends (Álvarez Nogal & Prados de la Escosura, 2013; Milanovic, 2016). Social mobility is a major topic in Historical Demography, proving its interaction with demographic behaviour (Dribe, Van Bavel, & Campbell, 2012; Schnore, 1961; Song, 2021). In this way, knowing the general context of inequality, and its evolution, appears to be determinant in any study of social stratification. However, the evolution of economic inequality in a long perspective is difficult due to the lack of direct sources (Alfani, 2015; Milanovic, Lindert, & Williamson, 2011; Piketty, 2015). The link between inequality and social mobility has been little explored from a historical view, and this will be one of our next research topics.

A long view analysis of the estimation of economic inequality was made possible thanks to the BHMD, which offers continuity over time (1451–1905) and space (the Barcelona area, the most populated area of Catalonia, also including the city of Barcelona). The fact that it is a fiscal source showing what each couple paid to get their marriage license meant that it was a measure of wealth. These taxes were imposed on the whole of society, which is not always the case with fiscal sources, one example being the tax exemption of the nobility throughout preindustrial periods (Pujadas-Mora & Brea-Martínez, 2020). Inequality in the Barcelona area was higher in preindustrial societies than in industrial ones, as recently stated by other authors (Alfani, 2015; Álvarez-Nogal & Prados de la Escosura, 2007; Milanovic, 2012, some of them already quoted) who disagree with the conclusions of the seminal work *Economic Growth and Income Inequality* (1955) by Simon Kuznets.

Three periods in the evolution of inequality in the Barcelona area are described. In the first, between 1481 and 1649, there was low inequality, which stabilized with the economic stagnation after the Catalan Civil War (1462–1472) and a whole series of mortality crises caused by different episodes of the plague. In the second, between 1650 and 1749, the levels of disparity increased significantly, coinciding with the transformation of the rural economy during the 17th century when new sharecropping contracts (*Rabassa Morta*) were established and the export of wine products increased. Inequality peaked around 1740 with the resurgence of the Catalan economy within the framework of an early proto-industrialization that took place in many rural areas, along with considerable population growth and the introduction of the first industrial establishments in Barcelona. In the third, between 1750 and 1880, there was a stabilization and subsequent decline in the disparity because of the War of Independence (1808–1814), followed by an upward trend that concurs with the date that traditionally marks the beginning of industrialization in Catalonia (1833). Inequality almost doubled between 1830 and 1880, driven by a significant increase in the levels of tax exemption (number of marriages receiving the marriage license by *Amore Dei*) among day laborers and workers.

As a result of decomposing the trends of economic inequality by social groups, it was observed that the nobility, liberal professionals, and merchants contributed positively to inequality, or in other words these groups had a proportionally greater weight in terms of wealth than population numbers. For their part, artisans, day laborer, and peasants contributed negatively by presenting a greater demographic weight than wealth. In fact, inequality in the Old Regime can be explained by a hierarchical society where the nobility concentrated much of the wealth, as expected, while from the second half of the 18th century onwards the groups that contributed to increasing inequality were those that worked in the factory system (day laborers, weavers, etc.), showing a clear proletarian effect. On analyzing the accumulation of wealth/income among the top 1% of the population along the five centuries covered by the BHMD, it was seen that by the end of the 17th century this group had gone from holding 10% of the total wealth to almost 17%, which was the highest point, decreasing again to 12% by 1880 (Pujadas-Mora & Brea-Martínez, 2020). In fact, the richest members of society were much richer in the preindustrial period, even though the development of capitalism allowed for greater levels of concentration. Socially, the nobles constituted more than 50% of the richest 1%, with their presence declining from the mid-17th century onwards to reach a figure of 10% at the end of the 18th century and during the 19th century, in a situation resembling what occurred in other European countries. The nobility was replaced by the large merchants and traders and the liberal professionals who by the end of the 19th century made up more than 80% of the top 1% in the Barcelona area.

Focusing the research on industrialization, it has been shown that labor market transformation and inequality were fundamental aspects in its consolidation (Brea-Martínez & Pujadas-Mora, 2018a). There was a massive incorporation of unskilled workers into industries from agriculture, the rural exodus caused by technical improvements in agricultural production and the development of the factory system (Martínez-Galarraga & Prat, 2016; Nadal, 1975; Sánchez Suárez, 1993). Using the BHMD, we show how all these processes

contributed to both the increased levels of economic inequality in the area of Barcelona along the 18th and 19th centuries and the transformation of the economic sectors. From 1780 onwards, there was an important decrease in the primary sector and a simultaneous increase in the secondary and tertiary sectors, which would advance the traditional starting date of industrialization in Catalonia. Moreover, between-sector economic inequality rose, as well as within the secondary sector (textile), due to a process of polarization. Conversely, the primary sector levelled-off in terms of disparity levels and the tertiary sector showed a declining pattern.

In general, the use of the BHMD for this specific research topic allowed us to introduce three fundamental innovations. First, using a marriage tax as a measure of wealth implies that the individuals are measured at the same time in the life cycle (in the same age range), making them more comparable to each other since the accumulation of wealth happens over time. Second, these tax records included everyone who married, even the nobles, who were otherwise almost universally tax exempt, and those exempt due to poverty. Thus, we have a cross-sectional view not usually provided by regular tax sources in Old Regime societies, which were subject tax fraud and tax evasion. And third, by merging the occupation or social status of the grooms together with the level of the marriage tax the economic concept of 'ability to pay' can be applied. The combination of the two magnitudes provides a more precise approach to the human capital of the individuals analyzed.

7 FINAL CONSIDERATIONS AND FUTURE AGENDA ON EXPANDING BHMD AND BALL DATABASES

The BHMD and the BALL provide proof of how artificial intelligence can facilitate the application of Handwritten Text Recognition (HTR) techniques to data collection of primary sources, which is key to building individual-level databases and reducing the time required for their construction. However, the current state-of-the-art of Handwriting Recognition means that some human intervention is still required, which explains the crowdsourcing and game-sourcing experiences implemented. Moreover, knowledge graph techniques have allowed the application of advanced record linkage, meaning that data complexity can be increased and its entire volume analyzed. In short, these approaches undoubtedly contribute to the so-called big data in terms of historical data.

Our future agenda is to continue building the BALL with the individual data of population registers/censuses and personal taxes for the 19th and 20th century from other Catalan counties. In December 2020, we initiated a crowdsourcing experience in the county of Maresme in the north of Barcelona. The forecast is that the same data for five new counties will be added in 2021. We estimate having 1.5 million individual observations by the end of the year. Besides, we plan to merge new textual sources to extend the individual socioeconomic information for the already compiled demographic data and document images to extract meaningful information to complement these demographic and economic data.

Future research in historical demography using the Barcelona Historical Marriage Database will examine:

- Assortative mating among French immigrant women in the great migratory wave received during the Hispanic Monarchy between the 16th and 18th centuries;
- Sibling similarities in social and occupational attainment in a long-term perspective from the 16th to the 19th century;
- Structural and strategic determinants of the formation of kin (isonymic) marriages;
- The association of residential segregation and socioeconomic inequality in 18th and 19th century Barcelona;
- The roles of politics and industrialization in the abandonment of the traditional marriage calendar;
- The diminishing influence of families on labor careers during the 19th and 20th centuries and the transition from stem families to nuclear families.

From the technological perspective, databases with the corresponding ground truth (images, transcriptions, named entity annotation) are being used to organize challenges and competitions in computer vision and document analysis and recognition. The databases will continue to be used for tasks like document object detection (layout segmentation), handwriting recognition, word spotting and information extraction in order

to advance in the automatic transcription of manuscript demographic sources. At the highest level, the representational models based on linked data structures will be a valid benchmark for testing algorithms on record linkage/link discovery. Another future line of research will be the transfer of the current algorithms to other types of archival data that contain complementary information for the databases, such as marriage advertisements in historical newspapers).

ACKNOWLEDGEMENTS

This work has been partially supported by the Spanish Science Ministry projects RTI2018-095533-B-I00, RTI2018-095645-B-C21, RYC-2014-16831, the Catalan project 2017-SGR-1783, the Catalan Ministry of Culture, the CERCA Program/Generalitat de Catalunya and the Culture Department of the Generalitat de Catalunya.

REFERENCES

- Alfani, G. (2015). Economic inequality in northwestern Italy: A long-term view (fourteenth to eighteenth centuries). *The Journal of Economic History*, 75(4), 1058–1096. Retrieved from https://EconPapers.repec.org/RePEc:cup:jechis:v:75:y:2015:i:04:p:1058-1096_00
- Almazán, J., Gordo, A., Fornés, A., & Valveny, E. (2014). Word spotting and recognition with embedded attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(12), 2552–2566. doi: [10.1109/TPAMI.2014.2339814](https://doi.org/10.1109/TPAMI.2014.2339814)
- Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies*, 1, 1–26. doi: [10.51964/hlcs9290](https://doi.org/10.51964/hlcs9290)
- Álvarez-Nogal, C., & Prados de la Escosura, L. (2007). The decline of Spain (1500–1850): Conjectural estimates. *European Review of Economic History*, 11(3), 319–366. Retrieved from https://EconPapers.repec.org/RePEc:cup:ereveh:v:11:y:2007:i:03:p:319-366_00
- Álvarez-Nogal, C., & Prados de la Escosura, L. (2013). The rise and fall of Spain (1270–1850). *The Economic History Review*, 66(1), 1–37. doi: [10.1111/j.1468-0289.2012.00656.x](https://doi.org/10.1111/j.1468-0289.2012.00656.x)
- Amengual-Bibiloni, M., & Pujadas-Mora, J. M. (2016). Orígens i destins de la immigració francesa a l'àrea de Barcelona (1481–1643). Aportacions a partir de la Barcelona Historical Marriage Database [Origins and destinations of French immigration in the area of Barcelona (1481–1643). Contributions from the Barcelona Historical Marriage Database]. *Manuscripts. Revista d'Història Moderna*, 34, 35–61. Retrieved from <https://dialnet.unirioja.es/servlet/articulo?codigo=5974891>
- Amengual-Bibiloni, M., & Pujadas-Mora, J. M. (2020). Cruce de caminos: Matrimonios de viudas y franceses en el área de Barcelona en los siglos XVI y XVII [Crossroads: Marriages between local widows and French migrants in the Barcelona area in the 16th and 17th centuries]. In Tovar Pulido, R. (Ed.), *De humilde e ilustre cuna: Retratos familiares de la España moderna (siglos XV-XIX)* (pp. 86–123). Évora : Publicações do Cidehus. doi: [10.4000/books.cidehus.10726](https://doi.org/10.4000/books.cidehus.10726)
- Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., & Patel-Schneider, P. F. (2003). *The description logic handbook: Theory, implementation and applications*. Cambridge: Cambridge University Press.
- Baucells, J. (2002). 'Esposalles' de l'arxiu de la catedral de Barcelona: Un fons documental únic (1451–1905) ['Marriage licenses' of the Barcelona cathedral's archive: A unique documentary collection (1451–1905)]. *Butlletí del Servei d'Arxius*, 35, 1–2.
- Berkner, L. K., & Mendels, F. F. (1976). Inheritance systems, family structure, and demographic patterns in Western Europe, 1700–1900. In C. Tilly, *Historical studies of changing fertility* (pp. 209–223). Princeton: Princeton University Press. doi: [10.1515/9781400871452-006](https://doi.org/10.1515/9781400871452-006)
- Bittles, A. H., & Black, M. L. (2010). Consanguineous marriage and human evolution. *Annual Review of Anthropology*, 39, 193–207. doi: [10.1146/annurev.anthro.012809.105051](https://doi.org/10.1146/annurev.anthro.012809.105051)
- Bloothoof, G. (1998). Assessment of systems for nominal retrieval and historical record linkage. *Computers and the Humanities*, 32(1), 39–56. Retrieved from <https://www.gerritbloothoof.nl/Publications/CHUM36.htm>

- Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26 (NIPS 2013)* (pp. 2787–2795). Retrieved from <https://papers.nips.cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf>
- Bourdieu, P. (1976). Marriage strategies as strategies of social reproduction. In R. Forster & O. Ranum (Eds.), *Family and society: Selections from the annales, économies, sociétés, civilisations* (pp. 117–144). Baltimore: Johns Hopkins University Press.
- Bourdieu, P. (2013). *Distinction: A social critique of the judgment of taste*. Abingdon: Routledge. (Original work published 1984)
- Bras, H., Kok, J., & Mandemakers, K. (2010). Sibship size and status attainment across contexts: Evidence from the Netherlands, 1840–1925. *Demographic Research*, 23(4), 73–104. doi: [10.4054/DemRes.2010.23.4](https://doi.org/10.4054/DemRes.2010.23.4)
- Brea-Martínez, G., & Pujadas-Mora, J. M. (2018a). Transformación y desigualdad económica en la industrialización en el área de Barcelona, 1715–1860 [Transformation and economic inequality in Industrialization in the area of Barcelona, 1715–1860]. *Revista de Historia Económica/ Journal of Iberian and Latin American Economic History*, 36(2), 241–273. doi: [10.1017/S0212610917000234](https://doi.org/10.1017/S0212610917000234)
- Brea-Martínez, G., & Pujadas-Mora, J. M. (2018b). Estimating long-term socioeconomic inequality in southern Europe: The Barcelona area, 1481–1880. *European Review of Economic History*, 23(4), 397–420. doi: [10.1093/ereh/hey017](https://doi.org/10.1093/ereh/hey017)
- Carbonell, M., Fornés, A., Villegas, M., & Lladós, J. (2020). A neural model for text localization, transcription and named entity recognition in full pages. *Pattern Recognition Letters*, 136, 219–227. doi: [10.1016/j.patrec.2020.05.001](https://doi.org/10.1016/j.patrec.2020.05.001)
- Carbonell, M., Riba, P., Villegas, M., Fornés, A., & Lladós, J. (2021). Named entity recognition and relation extraction with graph neural networks in semi structured documents. *2020 25th International Conference on Pattern Recognition (ICPR)*, 9622–9627. doi: [10.1109/ICPR48806.2021.9412669](https://doi.org/10.1109/ICPR48806.2021.9412669)
- Carreras Candi, F. (1913). Les obres de la Catedral de Barcelona 1298–1445 [The works of the Cathedral of Barcelona 1298–1445]. *Butlletí de la Reial Acadèmia de Bones Lletres de Barcelona*, 7(49), 22–30. Retrieved from <https://raco.cat/index.php/BoletinRABL/article/view/201731>
- Chacón Jiménez, F., & Hernández Franco, J. (Eds.) (1992). *Poder, familia y consanguinidad en la España del antiguo régimen* [Power, family and consanguinity in ancien régime Spain]. Madrid: Anthropos.
- Chen, J., Riba, P., Fornés, A., Mas, J., Lladós, J., & Pujadas-Mora, J. M. (2018). Word-Hunter: A gamesourcing experience to validate the transcription of historical manuscripts. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 528–533. doi: [10.1109/ICFHR-2018.2018.00098](https://doi.org/10.1109/ICFHR-2018.2018.00098)
- Das, S., Sial, H. A., Ma, K., Baldrich, R., Vanrell, M., & Samaras, D. (2020). Intrinsic decomposition of document images in-the-wild. *Proceedings of the British Machine Vision Conference (BMVC)*, 1–14. doi: [10.48550/arXiv.2011.14447](https://doi.org/10.48550/arXiv.2011.14447)
- Delille, G. (1994). Consanguinité proche en Italie du XVIe au XIXe siècle [Consanguinity in Italy from the sixteenth to the nineteenth century]. In Bonté, P. (Ed.), *Épouser au plus proche. Inceste, prohibitions et stratégies matrimoniales autour de la Méditerranée* (pp. 323–340). Paris: EHESS.
- Dribe, M., Van Bavel, J., & Campbell, C. (2012). Social mobility and demographic behavior: Long term perspectives. *Demographic Research*, 26, 173–190. doi: [10.4054/DemRes.2012.26.8](https://doi.org/10.4054/DemRes.2012.26.8)
- Ferrer-Alòs, Ll. (1983). Censals, vendes a carta de gràcia i endeutament pagès al Bages (s. XVIII) ['Censals', sales by letter of grace and peasant indebtedness in the Bages (18th century)]. *Estudis d'història agrària*, 4, 101–128. Retrieved from <https://raco.cat/index.php/EHA/article/view/99535>
- Ferrer-Alòs, Ll. (2014). Derechos de propiedad y mercado de la tierra en la Cataluña Vieja (s. XV–XIX). El caso de Artés (Bages) [Property rights and land market in Old Catalonia (15th–19th century). The case of Artés (Bages)]. *Historia agraria: Revista de agricultura e historia rural*, 62, 47–82. Retrieved from <https://ideas.repec.org/a/seh/journal/y2014i62maprilp47-82.html>
- Fornés, A., Lladós, J., Mas, J., Pujadas-Mora, J. M., & Cabré, A. (2014). A bimodal crowdsourcing platform for demographic historical manuscripts. *DATeCH '14: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 103–108. doi: [10.1145/2595188.2595199](https://doi.org/10.1145/2595188.2595199)

- Ganzeboom, H. B. G., de Graaf, P. M., & Treiman, D. J. (1991). A standard international socio-economic index of occupational status. *Social science research*, 21(1), 1–56. doi: [10.1016/0049-089X\(92\)90017-B](https://doi.org/10.1016/0049-089X(92)90017-B)
- García Espuche, A. (1998). *Un siglo decisivo: Barcelona y Cataluña, 1550–1640* [A decisive century: Barcelona and Catalonia, 1550–1640]. Barcelona: Alianza Editorial.
- García Pérez, M. S. (2007). El padrón municipal de habitantes: Origen, evolución y significado [Population register: Origin, evolution and significance]. *Hispania Nova: Revista de historia contemporánea*, 7, 1–5. Retrieved from <http://hispanianova.rediris.es/7/articulos/7a005.pdf>
- García Ruipérez, M. (2012). El empadronamiento municipal en España: Evolución legislativa y tipología documental [The population register in Spain: Legislative evolution and documentary typology]. *Documenta & Instrumenta*, 10, 45–86. doi: [10.5209/rev_DOCU.2012.v10.40485](https://doi.org/10.5209/rev_DOCU.2012.v10.40485)
- Gautam, B., Ramos-Terrades, O., Pujadas-Mora, J. M., & Valls, M. (2020). Knowledge graph based methods for record linkage. *Pattern Recognition Letters*, 136, 127–133. doi: [10.1016/j.patrec.2020.05.025](https://doi.org/10.1016/j.patrec.2020.05.025)
- Goiser, K., & Christen, P. (2006). Towards automated record linkage. *AusDM '06: Proceedings of the fifth Australasian conference on Data mining and analytics*, 61, 23–31.
- Gouesse, J. M. (1984). Mariages de proches parents (XVIe–XXe siècle). Esquisse d'une conjoncture [Marriages of close relatives (16th–20th century). Sketch of a conjuncture]. *Publications de l'École française de Rome, 90: Le modèle familial européen. Normes, déviations, contrôle du pouvoir. Actes des séminaires organisés par l'École française de Rome et l'Università di Roma* (pp. 31–61). Rome: École Française de Rome.
- Goyer, D. S., & Draaijer, G. E. (1992). *The handbook of national population censuses: Europe*. Santa Barbara: Greenwood.
- Guan, S., Jin, X., Jia, Y., Wang, Y., Shen, H., & Cheng, X. (2017). Self-learning and embedding based entity alignment. *IEEE International Conference on Big Knowledge (ICBK)*, 33–40. doi: [10.1109/ICBK.2017.15](https://doi.org/10.1109/ICBK.2017.15)
- Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). Record linkage — Methodology. In T. N. Herzog, F. J. Scheuren & W. E. Winkler, *Data quality and record linkage techniques* (pp. 81–92). New York: Springer. doi: [10.1007/0-387-69505-2_8](https://doi.org/10.1007/0-387-69505-2_8)
- Jordà, J. P., Ameijerías-Alonso, J., & Pujadas-Mora, J. M. (2018). Chronicle of an early demise, surname extinction in the fifteenth and the seventeenth centuries. *Historical Methods*, 51(3), 190–201. doi: [10.1080/01615440.2018.1462747](https://doi.org/10.1080/01615440.2018.1462747)
- Jordà, J. P., Pujadas-Mora, J. M., & Cabré, A. (2014). The footprint of migrations on surnames: Onomastic changes in the Barcelona area at the late Middle Ages (1451–1500). *Onoma*, 49, 105–136. doi: [10.2143/ONO.49.0.3285500](https://doi.org/10.2143/ONO.49.0.3285500)
- Jordà, J. P., Pujadas-Mora, J., M., & Cabré, A. (2016). Surnames and migrations: The Barcelona area (1451–1900). *Names and their Environment. Proceedings of the 25th International Congress of Onomastic*, 131–143. Retrieved from <https://ddd.uab.cat/record/174338>
- Kang, L., Riba, P., Rusiñol, P., Fornés, A., & Villegas, M. (2020b). Pay attention to what you read: Non-recurrent handwritten text-line recognition. *arXiv:2005.13044*. Retrieved from <https://arxiv.org/abs/2005.13044>
- Kang, L., Riba, P., Villegas, M., Fornés, A., & Rusiñol, M. (2020a). Candidate fusion: Integrating language modelling into a sequence-to-sequence handwritten word recognition architecture. *Pattern Recognition*, 112, 107790. doi: [10.1016/j.patcog.2020.107790](https://doi.org/10.1016/j.patcog.2020.107790)
- Kang, L., Rusiñol, M., Fornés, A., Riba, P., & Villegas, M. (2020c). Unsupervised adaptation for synthetic-to-real handwritten word recognition. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3491–3500. doi: [10.1109/WACV45572.2020.909339](https://doi.org/10.1109/WACV45572.2020.909339)
- Kleber, F., Diem, M., Dejean, H., Meunier, J.-L., & Lang, E. (2018). Matching table structures of historical register books using association graphs. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 217–222. doi: [10.1109/ICFHR-2018.2018.00046](https://doi.org/10.1109/ICFHR-2018.2018.00046)
- Kocka, J. (1984). Family and class formation: Intergenerational mobility and marriage patterns in nineteenth-century Westphalian towns. *Journal of Social History*, 17(3), 411–433. Retrieved from <https://www.jstor.org/stable/3787212>
- Kuznets, S. (1955). Economic growth and income inequality. *The American Economic Review*, 45(1), 1–28. Retrieved from <https://www.jstor.org/stable/1811581>
- Lambert, P. S., Zijdeman, R. L., van Leeuwen, M. H. D., Maas, I., & Prandy, K. (2013). The construction of HISCAM: A stratification scale based on social interactions for historical comparative research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 46(2), 77–89. doi: [10.1080/01615440.2012.715569](https://doi.org/10.1080/01615440.2012.715569)

- Laslett, P. (1983). Family and household as work group and kin group: Areas of traditional Europe compared. In R. Wall (Ed.), *Family forms in historic Europe* (pp. 513–564). Cambridge: Cambridge University Press. doi: [10.1017/CBO9780511897535.018](https://doi.org/10.1017/CBO9780511897535.018)
- Lin, Y., Liu, Z., & Sun, M. (2016). Knowledge representation learning with entities, attributes and relations. In S. Kambhampati (Ed.), *Proceedings of the twenty-fifth International Joint Conference on Artificial Intelligence (IJCAI-16)* (pp. 2866–2872). Palo Alto: AAAI Press/International Joint Conferences on Artificial Intelligence. Retrieved from <https://www.ijcai.org/Proceedings/16/Papers/407.pdf>
- Lladós, J., Rusinol, M., Fornés, A., Fernández, D., & Dutta, A. (2012). On the influence of word representations for handwritten word spotting in historical documents. *International Journal of Pattern Recognition and Artificial Intelligence*, 26(5), 1263002. doi: [10.1142/S0218001412630025](https://doi.org/10.1142/S0218001412630025)
- Low, B. S., & Clarke, A. L. (1991). Family patterns in nineteenth-century Sweden: Impact of occupational status and landownership. *Journal of Family History*, 16(2), 117–138. doi: [10.1177/036319909101600202](https://doi.org/10.1177/036319909101600202)
- Lucassen, J., & Lucassen, L. (2009). The mobility transition revisited, 1500-1900: What the case of Europe can offer to global history. *Journal of Global History*, 4(3), 347–377. doi: [10.1017/S174002280999012X](https://doi.org/10.1017/S174002280999012X)
- Marín Corbera, M. (2010). La gestación del Documento Nacional de Identidad: Un proyecto de control totalitario para la España Franquista [The gestation of the National Identity Card: A totalitarian control project for Francoist Spain]. In C. Navajas Zubeldia & D. Iturriaga Barco (Eds.), *Novísima: II Congreso Internacional de Historia de Nuestro Tiempo* (pp. 323–338). Retrieved from <https://dialnet.unirioja.es/servlet/articulo?codigo=3313002>
- Martín Niño, J. (1972). *La hacienda española y la revolución de 1868* [The Spanish treasury and the 1868 revolution]. Madrid: Instituto de Estudios Fiscales.
- Martínez-Galarraga, J., & Prat, M. (2016). Wages, prices, and technology in early Catalan industrialization. *The Economic History Review*, 69(2), 548–574. doi: [10.1111/ehr.12127](https://doi.org/10.1111/ehr.12127)
- Mas, J., Fornés, A., & Lladós, J. (2016). An interactive transcription system of census records using word-spotting based information transfer. *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 54–59. doi: [10.1109/DAS.2016.47](https://doi.org/10.1109/DAS.2016.47)
- Melis Maynar, F. (2019). Distribución personal y provincial de la renta en 1926 según el impuesto de cédulas personales [Personal and provincial distribution of income in 1926 according to the personal taxes]. *Papeles de trabajo del Instituto de Estudios Fiscales. Serie economía*, 3, 1–44. Retrieved from <https://dialnet.unirioja.es/servlet/articulo?codigo=7206051>
- Milanovic, B. (2012). *Global income inequality by the numbers: In history and now* (Policy Research Working Paper; No. 6259). Washington, DC: World Bank. Retrieved from <https://openknowledge.worldbank.org/handle/10986/12117>
- Milanovic, B. (2016). *Global inequality: A new approach for the age of globalization*. Cambridge: Harvard University Press.
- Milanovic, B., Lindert, P. H., & Williamson, J. G. (2011). Pre-industrial inequality. *The Economic Journal*, 121(551), 255–272. doi: [10.1111/j.1468-0297.2010.02403.x](https://doi.org/10.1111/j.1468-0297.2010.02403.x)
- Moreu Rey, E. (1959). *Els immigrants francesos a Barcelona (segles XVI al XVIII)* [French immigrants in Barcelona (16th to 18th centuries)] (Vol. 20). Barcelona: Institut d'Estudis Catalans.
- Nadal, J. (1975). *El fracaso de la revolución industrial en España, 1814–1913* [The failure of the industrial revolution in Spain, 1814–1913]. Barcelona: Ariel.
- Nadal, J., & Giralt, E. (1960). *La población catalana de 1553 a 1717: L'immigración francesa et les autres facteurs de son développement* [The Catalan population from 1553 to 1717: The French emigration and the others factors of its development] (Vol. 3). Paris: S.E.V.P.E.N.
- Pérez Hernández, E. (2020). *The distribution of income in Madrid over the first half of the 20th century: An estimation using rental values* (Master thesis). Universidad Carlos III, Madrid.
- Peytaví Deixona, J. (2010). *Antroponimia, poblament i immigració a la Catalunya moderna. L'exemple dels comtats de Rosselló i Cerdanya (segles XVI-XVIII)* [Anthroponymy, settlement and immigration in modern Catalonia. The example of the counties of Rosselló and Cerdanya (16th–18th centuries)]. Barcelona: IEC. Retrieved from <https://publicacions.iec.cat/repository/pdf/00000233/00000077.pdf>
- Piketty, T. (2015). *The economics of inequality*. Cambridge: Harvard University Press.
- Pujadas-Mora, J. M., & Brea-Martínez, G. (2020). Five centuries of inequality and socioeconomic transformation in the Barcelona area, 1451–1880. *Perspectives Demogràfiques*, 18, 1–4. doi: [10.46710/ced.pdf.eng.18](https://doi.org/10.46710/ced.pdf.eng.18)

- Pujadas-Mora, J. M., Brea-Martínez, G., Jordà, J.P., & Cabré, A. (2018). The apple never falls far from the tree: Siblings and intergenerational transmission among farmers and artisans in the Barcelona area in the sixteenth and seventeenth centuries. *The History of the Family*, 23(4), 533–567. doi: [10.1080/1081602X.2018.1426483](https://doi.org/10.1080/1081602X.2018.1426483)
- Pujadas-Mora, J. M., Fornés, A., Lladós, J., Brea-Martínez, G., & Valls-Fígols, M. (2019). The Baix Llobregat (BALL) demographic database, between historical demography and computer vision (nineteenth–twentieth centuries). In E. Glavatskaya, G. Thorvaldsen, Georg F., & M. Szoltysek (Eds.), *Nominative data in demographic research in the East and the West* (pp. 29–61). Ekaterinburg: Ural University Press. doi: [10.15826/B978-5-7996-2656-3.03](https://doi.org/10.15826/B978-5-7996-2656-3.03)
- Pujadas-Mora, J. M., Fornés Bosquerra, A., Lladós, J., & Cabré, A. (2016). Bridging the gap between historical demography and computing: Tools for computer-assisted transcription and analysis of demographic sources. In K. Matthijs, S. Hin, J. Kok, & H. Matsuo, *The future of historical demography: Upside down and inside out* (pp. 222–226). Leuven/Den Haag: Acco. Retrieved from <https://soc.kuleuven.be/ceso/fapos/publications/the-future-of-historical-demography-upside-down-and-inside-out>
- Pujadas-Mora, J. M., González-Murciano, C., & Cabré, A. (2018). Fenomen rodstvennykh brakov v Katalonii v XIX v. (po materialam istoricheskoy bazy dannykh Barselony) [Kin marriages in the 19th century Catalonia. With reference to findings from the Barcelona historical marriage database]. *Izvestia. Ural Federal University Journal. Series 2. Humanities and Arts*, 20(4(181)), 27–45. doi: [10.15826/izv2.2018.20.4.064](https://doi.org/10.15826/izv2.2018.20.4.064)
- Pujadas-Mora, J. M., Romeo-Marín, J., & Villar, C. (2014). Propuestas metodológicas para la aplicación de HISCO en el caso de Cataluña, siglos XV-XX [Methodological proposals for the application of HISCO to the case of Catalonia, 15th–20th centuries]. *Revista de Demografia Histórica*, 32(1), 181–219. Retrieved from <https://dialnet.unirioja.es/servlet/articulo?codigo=5161081>
- Reher, D. S., & Valero Lobo, A. (1995). *Fuentes de información demográfica en España* [Sources of demographic information in Spain] (Vol. 13). Madrid: CIS.
- Rubio Vela, A., & Rodrigo Lizondo, M. (1997). *Antroponímia valenciana del segle XIV: Nòmnes de la ciutat de València (1368–69 i 1373)* [Valencian anthroponymy of the 14th century: Name list of the city of Valencia (1368–69 and 1373)]. Valencia: Universitat de València.
- Salas-Vives, P., & Pujadas-Mora, J. M. (2021). Bottom-up nation-building: National censuses and local administration in nineteenth-century Spain. *Journal of Historical Sociology*, 34(2), 287–304. doi: [10.1111/johs.12323](https://doi.org/10.1111/johs.12323)
- Sánchez Suárez, A. (1993). 1993 Les activitats econòmiques a Barcelona (1717–1833) [Economic activities in Barcelona (1717–1833)]. In J. Sobrequés i Callicó (Ed.), *Història de Barcelona. Vol. 5: El desplegament de la ciutat manufacturera (1714–1833)* (pp. 217–265). Barcelona: Enciclopedia Catalana.
- Schlomer, G. L., Ellis, B. J., & Garber, J. (2010). Mother–child conflict and sibling relatedness: A test of hypotheses from parent–offspring conflict theory. *Journal of Research on Adolescence*, 20(2), 287–306. doi: [10.1111/j.1532-7795.2010.00641.x](https://doi.org/10.1111/j.1532-7795.2010.00641.x)
- Schneider, E. W. (1973). Course modularization applied: The interface system and its implications for sequence control and data analysis. *Association for the Development of Instructional Systems (ADIS)*, 1–17. Retrieved from <https://files.eric.ed.gov/fulltext/ED088424.pdf>
- Schnore, L. F. (1961). Social mobility in demographic perspective. *American Sociological Review*, 26(3), 407–423. doi: [10.2307/2090668](https://doi.org/10.2307/2090668)
- Schürer, K. (2007). Focus: Creating a nationally representative individual and household sample for Great Britain, 1851 to 1901: the Victorian Panel Study (VPS). *Historical Social Research/ Historische Sozialforschung*, 32(2), 211–331. doi: [10.12759/hsr.32.2007.2.211-331](https://doi.org/10.12759/hsr.32.2007.2.211-331)
- Shenk, M. K., Borgerhoff Mulder, M., Beise, J., Clark, G., Irons, W. G., Leonetti, D., ... Piraino, P. (2010). Intergenerational wealth transmission among agriculturalists: Foundations of agrarian inequality. *Current Anthropology*, 51(1), 65–83. doi: [10.1086/648658](https://doi.org/10.1086/648658)
- Singhal, A. (2012). Introducing the Knowledge Graph: Things, not strings. *Official Google Blog*. Retrieved from <https://blog.google/products/search/introducing-knowledge-graph-things-not/>
- Song, X. (2021). Multigenerational social mobility: A demographic approach. *Sociological Methodology*, 51(1), 1–43. doi: [10.1177/0081175020973054](https://doi.org/10.1177/0081175020973054)
- Sutter, J. (1968). Fréquence de l'endogamie et ses facteurs au XIXe siècle [Frequency of endogamy and its factors in 19th century]. *Population*, 23(2), 303–324. doi: [10.2307/1527490](https://doi.org/10.2307/1527490)
- Toledo, J. I., Carbonell, M., Fornés, A., & Lladós, J. (2019). Information extraction from historical handwritten document images with a context-aware neural model. *Pattern Recognition*, 86, 27–36. doi: [10.1016/j.patcog.2018.08.020](https://doi.org/10.1016/j.patcog.2018.08.020)

- Toledo, J. I., Dey, S., Fornés, A., & Lladós, J. (2017). Handwriting recognition by attribute embedding and recurrent neural networks. *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, 1038–1043. doi: [10.1109/ICDAR.2017.172](https://doi.org/10.1109/ICDAR.2017.172)
- Toledo, J. I., Sudholt, S., Fornés, A., Cucurull, J., Fink, G. A., & Lladós, J. (2016). Handwritten word image categorization with convolutional neural networks and spatial pyramid pooling. In A. Robles-Kelly, M. Loog, B. Biggio, F. Escolano & R. Wilson (Eds.) *Structural, Syntactic, and Statistical Pattern Recognition. S+SSPR 2016. Lecture Notes in Computer Science* (Vol. 10029) (pp. 543–552). Cham: Springer. doi: [10.1007/978-3-319-49055-7_48](https://doi.org/10.1007/978-3-319-49055-7_48)
- Torras, J. (1998). Small towns, craft guilds and proto-industry in Spain. *Jahrbuch für Wirtschaftsgeschichte/ Economic History Yearbook*, 39(2), 79–96. doi: [10.1524/jbwg.1998.39.2.79](https://doi.org/10.1524/jbwg.1998.39.2.79)
- Trivers, R. L. (1974). Parent-offspring conflict. *American Zoologist*, 14(1), 249–264. doi: [10.1093/icb/14.1.249](https://doi.org/10.1093/icb/14.1.249)
- Van de Putte, B., & Miles, A. (2005). A social classification scheme for historical occupational data. *Historical Methods*, 38(2), 61–94. doi: [10.3200/HMTS.38.2.61-94](https://doi.org/10.3200/HMTS.38.2.61-94)
- van Leeuwen, M. H. D., & Maas, I. (2011). *HISCLASS: A historical International Social Class Scheme*. Leuven: University Press.
- van Leeuwen, M. H. D., Maas, I., & Miles, A. (2002). *HISCO: Historical International Standard Classification of Occupations*. Leuven: University Press.
- Vilar, P. (1986). *Catalunya dins l'Espanya moderna: III. Les transformacions agràries del segle XVIII català* [Catalonia in modern Spain: III. The agrarian transformations of the Catalan 18th century]. Barcelona: Edicions 62.
- Villavicencio, F., Jordà, J. P., & Pujadas-Mora, J. M. (2015). Reconstructing lifespans through historical marriage records of Barcelona from the sixteenth and seventeenth centuries. In G. Bloothoof, P. Christen, K. Mandemakers & M. Schraagen (Eds.), *Population Reconstruction* (pp. 199–216). Cham: Springer. doi: [Pu](https://doi.org/10.1007/978-3-319-18870-0_10)
- Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014). Knowledge graph embedding by translating on hyperplanes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1), 1112–1119. Retrieved from <https://ojs.aaai.org/index.php/AAAI/article/view/8870>

APPENDIX

Appendix Table 1 *Number of inhabitants and households by town included in BALL database, 19th–20th centuries (1822–1965)*

Population registers and national censuses

Sant Feliu de Llobregat																			
	1828	1839	1857	1878	1881	1889	1906	1910	1915	1920	1924	1930	1936	1940	1945	1950	1955	1960	1965
Total inhabitants	2,209	2,233	2,471	2,747	3,002	3,118	3,606	3,807	4,329	4,352	5,569	6,383	7,020	6,720	6,977	7,327	7,529	10,201	12,945
Total households	426	432	533	673	641	645	804	867	937	923	1,048	1,162	1,458	1,678	1,702	1,812	2,153	2,784	3,742
Persons by household	5.19	5.17	4.64	4.08	4.68	4.83	4.49	4.39	4.62	4.72	5.31	5.49	4.81	4.00	4.10	4.04	3.50	3.66	3.46

Santa Coloma de Cervelló						
	1901	1924	1936	1940	1945	1950
Total inhabitants	542	1,132	1,311	1,218	1,167	1,222
Total households	127	260	325	329	307	308
Persons by household	4.27	4.35	4.03	3.70	3.80	3.97

Sant Vicenç dels Horts										
	1857	1860	1921	1925	1930	1935	1940	1946	1950	1955
Total inhabitants	1,772	1,731	2,097	1,985	2,950	3,129	2,980	3,014	3,323	3,717
Total households	357	350	535	409	783	716	763	791	1,030	1,168
Persons by household	4.96	4.95	3.92	4.85	3.77	4.37	3.91	3.81	3.23	3.18

Collbató																			
	1842	1845	1857	1860	1863	1866	1867	1868	1869	1870	1871	1872	1875	1880	1887	1889	1892	1896	1897
Total inhabitants	794	746	865	859	840	828	790	765	785	802	798	828	795	791	781	812	743	686	679
Total households	109	115	140	152	154	145	139	142	148	153	160	157	149	168	172	170	154	147	148
Persons by household	7.28	6.49	6.18	5.65	5.45	5.71	5.68	5.39	5.30	5.24	4.00	5.27	5.34	4.71	4.54	4.78	4.82	4.67	4.59

Collbató (continued)											
	1900	1905	1910	1916	1920	1924	1930	1936	1940	1945	1950
Total inhabitants	703	665	809	796	560	754	531	525	474	425	418
Total households	182	144	162	166	141	170	138	138	162	125	147
Persons by household	3.86	4.62	4.99	4.80	3.97	4.44	3.85	3.80	2.93	3.40	2.84

Population registers and national censuses (continued)

Castellví de Rosanes							
	1866	1924	1930	1936	1940	1945	1950
Total inhabitants	353	283	279	273	281	242	268
Total households	63	61	64	62	73	50	60
Persons by household	5.60	4.64	4.36	4.40	3.85	4.84	4.47

Molins de Rei														
	1852	1871	1872	1899	1905	1911	1915	1920	1924	1925	1930	1940	1950	1955
Total inhabitants	1,339	3,387	**	2,896	3,003	3,354	3,383	3,842	4,917	**	6,469	7,622	**	**
Total households	259	502		856	717	763	817	982	1,093		1,845	2,212		
Persons by household	5.17	6.75		3.38	4.19	4.40	4.14	3.91	4.50		3.51	3.45		

El Papiol														
	1857	1875	1885	1889	1890	1895	1900	1916	1921	1924	1936	1940	1945	1950
Total inhabitants	1,100	**	**	1,022	**	**	912	**	**	**	**	**	**	**
Total households	198			190			183							
Persons by household	5.56			5.38			4.98							

Martorell							
	1842	1916	1924	1940	1945	1950	1955
Total inhabitants	**	3,822	**	6,113	6,214	**	**
Total households		906		1,940	1,936		
Persons by household		4.22		3.15	3.21		

Olesa de Montserrat								
	1860	1872	1875	1910	1915	1920	1924	1930
Total inhabitants	3,176	3,148	2,810	3,850	3,734	4,060	4,369	5,647
Total households	621	690	620	927	824	951	949	1,330
Persons by household	5.11	4.56	4.53	4.15	4.53	4.27	4.60	4.25

Population registers and national censuses (continued)

Pallejà							
	1827	1857	1889	1910	1924	1945	1965
Total inhabitants	642	838	718	653	800	998	3,127
Total households	111	178	166	169	178	237	705
Persons by household	5.78	4.71	4.33	3.86	4.49	4.21	4.44

Abrera													
	1850	1855	1857	1865	1875	1889	1906	1912	1920	1930	1936	1940	1945
Total inhabitants	**	431	**	**	**	875	838	**	835	**	**	**	**
Total households		103				233	175		226				
Persons by household		4.18				3.76	4.79		3.69				

Torrelles de Llobregat																			
	1842	1845	1850	1852	1857	1860	1862	1863	1864	1865	1866	1867	1887	1889	1890	1895	1897	1900	1906
Total inhabitants	465	388	533	**	494	496	457	**	**	**	**	**	675	**	705	697	683	693	715
Total households	72	74	103		96	105	98						140		129	120	122	153	160
Persons by household	6.46	5.24	5.17		5.15	4.72	4.66						4.82		5.47	5.81	5.60	4.53	4.47

Torrelles de Llobregat (continued)									
	1910	1917	1920	1924	1930	1936	1940	1945	1950
Total inhabitants	726	**	755	767	817	1,573	762	733	780
Total households	173		180	156	192	326	198	191	187
Persons by household	4.20		4.19	4.92	4.26	4.83	3.85	3.84	4.17

Begues										
	1857	1860	1871	1881	1900	1905	1910	1915	1924	1928
Total inhabitants	**	858	902	955	**	**	**	**	**	**
Total households		158	176	193						
Persons by household		5.43	5.13	4.95						

Population registers and national censuses (continued)

Corbera de Llobregat																		
	1857	1860	1877	1880	1888	1892	1897–98	1900	1916	1920	1921	1923	1924	1930	1936	1940	1945	1950
Total inhabitants	**	869	**	981	1,022	981	**	964	1,092	1,197	**	**	1,301	1,247	1,271	1,175	**	**
Total households		180		276	234	274		230	240	232			259	272	282	278		
Persons by household		4.83		3.55	4.37	3.58		4.19	4.55	5.16			5.02	4.58	4.51	4.23		

** Under construction

Personal taxes

San Feliu de Llobregat																			
	1868	1869	1882	1883	1884	1885	1886–89	1890	1891	1892	1907	1909	1911	1912	1913–15	1916	1917	1918	1919
Personal taxes	438	**	1,185	426	352	393	**	986	903	**	1,891	1,849	**	2,057	**	2,861	**	2,822	**

San Feliu de Llobregat (continued)

	1920	1921–23	1924	1925–28	1929	1930–41
Personal taxes	2,894	**	2,673	**	3,848	**

Santa Coloma de Cervelló

	1910	1913–15	1916	1917–19	1920	1921–23	1924	1925	1927–28	1929	1931–34	1935	1937	1941
Personal taxes	221	**	298	**	308	**	612	**	**	644	**	934	**	**

Sant Vicenç dels Horts

	1879	1880	1881
Personal taxes	325	303	332

Collbató

	1878–79	1879–80	1880–81	1881–82	1883–84	1884–85	1885–86
Personal taxes	202	168	176	218	488	486	482

Personal taxes (continued)

Torrelles de Llobregat												
	1882–86	1887	1888–91	1906	1907–08	1911	1912–19	1920	1921–23	1924	1925–29	1930
Personal taxes	**	327	**	460	**	492	**	523	**	530	**	569

Corbera de Llobregat							
	1881–84	1888	1892–93	1920–21	1924	1930	1936
Personal taxes	**	674	**	**	792	**	**

Martorell								
	1911	1916	1919	1923–24	1927	1931	1935	1940
Personal taxes	**	2,780	**	**	**	**	**	**

Abdera				
	1921	1922–24	1928–32	1934
Personal taxes	626	**	**	**

** Under construction

SLAVERY IN SURINAME

A Reconstruction of Life Courses, 1830–1863

Coen W. van Galen	Radboud University Nijmegen
Rick J. Mourits	International Institute of Social History, Amsterdam & Radboud University Nijmegen
Matthias Rosenbaum-Feldbrügge	Radboud University Nijmegen
Maartje A.B.	Radboud University Nijmegen
Jasmijn Janssen	Radboud University Nijmegen
Björn Quanjer	Radboud University Nijmegen
Thunnis van Oort	Radboud University Nijmegen
Jan Kok	Radboud University Nijmegen

ABSTRACT

The *slavenregisters* or slave registers of Suriname offer a unique perspective on the social and demographic history of a people in bondage. Thanks to a citizen science project, the archival sources were transcribed in 2017 by hundreds of volunteers. The transcriptions were used to create a longitudinal database of more than 90,000 enslaved persons. This paper describes the sources, data entry, and cleaning to create a standardized database as well as the matching needed to construct life courses. We discuss the best practices we have learned along the way. Finally, it offers prospects for research and expansion of the database to other population sources and areas.

Keywords: Slavery, Suriname, Life courses, Citizen science, Record linkage

DOI article: <https://doi.org/10.51964/hlcs15619>

© 2023, van Galen, Mourits, Rosenbaum-Feldbrügge, A.B., Janssen, Quanjer, van Oort, Kok
This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

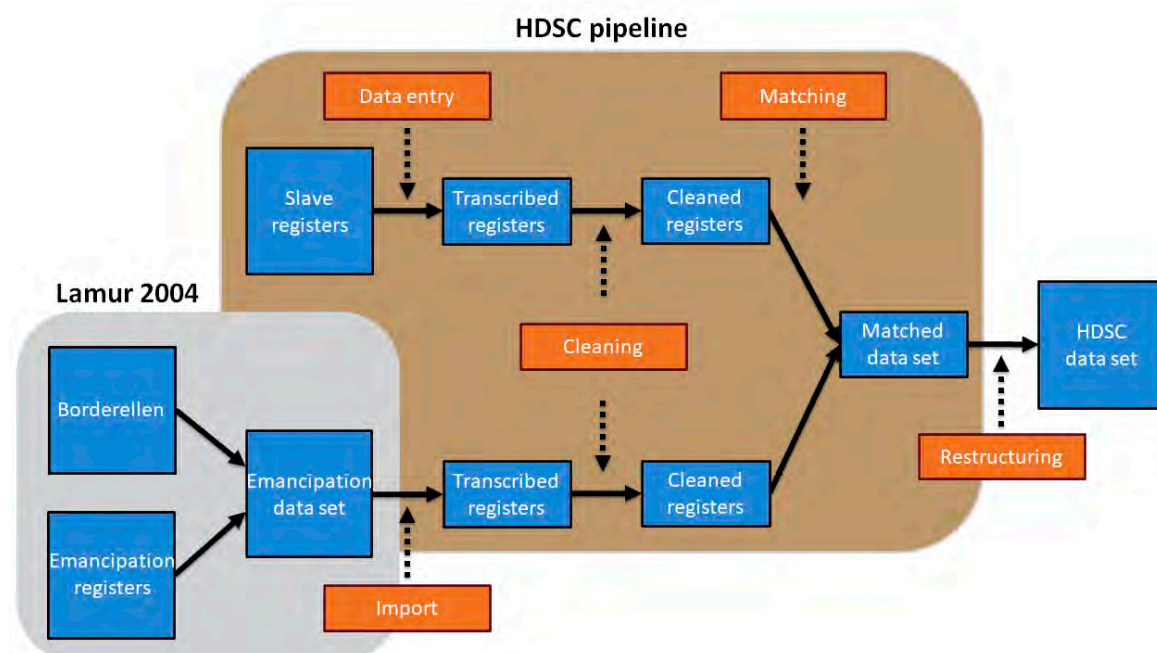
In recent decades a growing corpus of population databases enables large-scale population research by reconstructing the life courses of sizable groups of people in the 19th and 20th centuries. However, these databases are almost all concentrated on Europe, the Western Offshoots and East Asia (see for a recent overview [Mandemakers, 2023](#)). Excellent research has been done on the demography of enslaved populations. Examples are Higman (1976, 1984) and Meridith John (1988) on the Caribbean; Menard (1975) and Margo and Steckel (1982) on North America; and Lamur (1981, 1987), van Stipriaan (1993, pp. 310–346) and Everaert (1999, 2013) on Suriname. Yet, the attention of historians of colonialism and slavery has largely shifted towards their cultural impact. Furthermore, recent large databases that have been created on slavery and the slave trade tend to focus on experiences or on specific events, not on the life courses of the individuals involved (Eltis, Halbert, & Misevich, 2018; [Enslaved.org, 2018](#); [Hall & Draper, 2019](#); [Lovejoy, 2015](#); [Midlo Hall, Hawthorne, & Mitchell, 2019](#)). The Historical Database of Suriname and the Caribbean (HDSC) is an attempt to bridge both types of databases.

The aim of the HDSC is to reconstruct the entire population of the former Dutch colonies Suriname and the Netherlands Antilles between 1830 and 1950. Reconstructing the lives of Surinamese enslaved people between 1830 and 1863 is the first step of this ambitious project. This article discusses how the first part of this database was built, in particular using the slave registers, which were linked to the Emancipation database (Lamur, 2004). A large part of the population (about 70%) of Suriname was enslaved until the abolition of slavery on July 1, 1863.

The slave registers were an attempt by the Dutch government in the early 19th century to create a complete accounting of enslaved people. The intention was to track every individual through time from one slave owner to another as long as she or he was in slavery. This means that, for Suriname, there is information not only on enslaved people on some large plantations, but also on enslaved people who worked in households, in workshops and even in churches. Such comprehensive material is available for almost no other colonial context and certainly not over such a long period.

On top of that, people released from slavery were immediately entered in the civil registry in the Dutch colonies. This included both persons freed when slavery ended in Suriname in 1863, and manumitted persons who were freed before the end of slavery. That means we can continue to track them after the end of slavery. Because surnames were compulsory in the civil registers and enslaved persons were not allowed to have surnames, in 1863 all of them received new surnames and often new first names as well. Registers of these new names were created in 1863. Lamur (2004) used these registers with lists of all enslaved persons from 1862 to create his Emancipation database. We have used this database with the permission of the author. Figure 1 shows the Lamur 2004 and HDSC pipeline.

Figure 1 *Flow chart of the construction of the Historical Database of Suriname and the Caribbean*



In this paper, we will first discuss the institutional setting of the Historical Database of Suriname and the Caribbean. Furthermore, we discuss the origins and the peculiarities of the slave registers. Then we will describe step by step the way the data were collected, cleaned, and matched. Finally, we discuss the output formats and the possibilities and limitations of the reconstructed life courses of enslaved people for research. We end with some concluding remarks.

Figure 1 gives an overview of the step-by-step process of transforming the slave registers, borderel, and emancipation registers into the dataset. Sources, the cleaned databases, and matched registrations/releases are all stored separately (see e.g., [Mandemakers & Dillon, 2004](#)).

2 THE HISTORICAL DATABASE OF SURINAME AND THE CARIBBEAN: INSTITUTIONAL SETTING

The Historical Database of Suriname and the Caribbean (HDSC) Foundation was established in order to publish databases of the population of Suriname and the Dutch Caribbean. Its aim is to reconstruct the entire population of the former Dutch colonies Suriname and the Dutch Caribbean between 1830 and 1950. These databases are both constructed for historical and demographic research and for a general audience, more specific for genealogy, education and cultural projects. For researchers, charting mortality, family formation and migration of as many individuals as possible offers the unique opportunity to study demographic and social processes in colonial, tropical societies.

The HDSC is founded in the Netherlands and embedded in the Radboud University in Nijmegen, where the HDSC technical staff works. The Members of the Board consist of historians and archivists from Suriname, Curacao and the Netherlands. The HDSC functions as a network, in which it participates with different partners in subprojects. Partners of the HDSC are, among others, the National Archives of Suriname, Curacao and the Netherlands and the Anton de Kom University of Suriname. The HDSC's core principle is the full and open availability of data for both scholarly research and the general public. This principle addresses the disparity in data accessibility between Dutch and Caribbean researchers. It is also a recognition that the data does not belong to the HDSC. It is a shared heritage of the peoples of Suriname, the Dutch Caribbean and the Netherlands.

In 2017, the HDSC started as a Surinamese-Dutch crowdfunding and citizen science project called "Maak de Surinaamse slavenregisters openbaar" [Make the Surinamese slave registers public]. The successful crowdfunding campaign, together with additional funding by the Prins Bernhard Cultuurfonds, Stichting Democratie en Media and CLARIAH, made it possible to digitize the slave registers by the National Archives of Suriname and to fund the use of the online citizen science platform Vele Handen (www.velehanden.nl). Although commercial, this platform was chosen, because it is also used by many Dutch archives and many volunteers were already familiar with Vele Handen.

Online transcription allowed participants from all over the world to join and transcribe the slave registers from home. A citizen science project was preferred in order to involve a broad audience and news media in the project, and to create public awareness of the role of slavery in Dutch history. Moreover, it would limit transcription costs in both time and money, and consequently make the data rather quickly available to a large audience ([Bonney et al., 2009](#); [Cohn, 2008](#); [Franzoni & Sauermann, 2014](#); [Irwin, 1995](#)).

Considerable effort was made to involve a wide audience in the project, with the intention to recruit many volunteers so that the pace of processing data would remain high. In this way, we hoped to avoid the usual pitfalls of citizen science projects, such as a difficulty to attract volunteers, dwindling interest during the project or concerns about the quality of the volunteers ([Bone et al., 2012](#); [Crall et al., 2017](#); [Sauermann & Franzoni, 2015](#); [van Galen, 2019](#)). More than 600 Dutch and Surinamese volunteers were recruited through social media and news broadcasts. A core group of 328 participants stayed during the entire project and became experts in reading the Surinamese slave registers. Thanks to this large number of participants, the 17,500 scans of the Surinamese slave registers were transcribed within four months, resulting in publication of the database on the websites of the Dutch and Surinamese National Archives in 2018–2019 ([van Galen, A.B., Mourits, & Rosenbaum-Feldbrügge, 2019](#)).

The datasets built by the HDSC can be used independently, but also combined to follow groups or individuals through time, even over multiple generations. Furthermore, the HDSC will release auxiliary datasets that are used to clean variables in the slave and emancipation registers, such as an overview of existing plantation names and available occupational titles. Additional datasets, such as the Suriname plantation dataset (Rosenbaum-Feldbrügge, van Galen, & Swaters, 2023), will help to further enhance the research capabilities of this database. Standardization of occupations will require the development of a specific thesaurus of unfree labour that will also be applicable elsewhere in the Caribbean for historical populations in slavery, as the labour structure, occupational titles, and meaning of occupational titles can differ considerably from western populations.

In addition to improving the current database of the Surinamese slave and emancipation registers, the HDSC will develop in two directions. The first direction is to expand the database through time by linking it to the civil registry so that the descendants of former enslaved people, as well as those of former slave owners, can be tracked down through the generations (Rosenbaum-Feldbrügge, Mourits, van Oort, & van Galen, 2023). Due to privacy regulations, the limit has been set at 1950 for now, making long-term processes over five generations visible. The second direction is to include other archival sources, such as the manumission registers and the registers of contract workers from China, India, Indonesia and the Caribbean, people who came to Suriname to replace the enslaved workers on the plantations. Collectively, these sources will provide greater insight into the changing multicultural colonial society that was Suriname.

The HDSC will also develop into a much broader database by including the population of the Dutch Caribbean. A database of the slave registers and emancipation records of Curacao, has already been published (Langeveld, van Galen, Quanjer, & Paul, 2020). The data on Curacao will also be expanded to include information on the population up to 1950. Publications on other islands such as St. Eustatius and Aruba are also published (Arends, Raaijmakers, Rosenbaum-Feldbrügge, & van Galen, 2023; Raaijmakers, & van Galen, 2023). However, societies in the Caribbean are not isolated and can only be properly studied by taking into account migration patterns that often ignored national boundaries. Hence, an inventory of relevant sources in areas outside the Dutch colonial context, such as the Danish Virgin Islands, the English and French Caribbean, the Dominican Republic and elsewhere, is being prepared.

3 SOURCES

The Surinamese slave registers were introduced in 1826 to counteract the illegal trans-Atlantic slave trade. This decision was made under pressure from the British government (van Galen & Hassankhan, 2018). During the Napoleonic wars, the British had occupied the Dutch colonies in the Americas. In 1815, the colony of Suriname (see Figure 2) was returned to the Dutch after the establishment of the Kingdom of the Netherlands, provided that the Netherlands accepted a ban on the trans-Atlantic slave trade which was already in force in the British colonies (Buddingh, 2022). As smuggling continued into the 1820s, the slave registers were introduced to curb the illegal international slave trade by registering all enslaved persons.

The Surinamese slave registers list the enslaved persons that were owned by plantations or private slave owners. Individuals owned by the colonial government were only registered after 1850. The books give information about the lives of approximately 90,000 enslaved individuals, 375 plantations, and over 3,500 private owners. Changes in ownership and vital events were duly noted, so that, at least in theory, the number of enslaved persons per owner was known for each point in time between 1830 and 1863. This contrasts with British slave registers in the region, which registered enslaved populations via a census every three years (Higman, 1984, pp. 6–15).

The slave registers of Suriname were kept up-to-date by a special civil servant in Paramaribo, who recorded each enslaved individual's personal information and changes in ownership (see Figure 3). Every so often, information that was still relevant was transcribed into new register books to keep the registration orderly. After an initial phase from 1826 onwards, four series of the Surinamese slave register came into being: 1830–1838, 1838–1848, 1848–1851, 1851–1863, referred to in this paper as series 1 to 4 (see Table 1).

Figure 2 Suriname around 1882



Source: *Kuyper & Heyse, ca. 1882.*

Note: Even if this map was produced after the period under consideration, it depicts the plantation areas that were active during the 19th century (marked in brown). It also shows the vast acreage that was never cultivated for commercial plantations.

Figure 3 Example of slave registers series 2 (1838–1848)

No Fol. 4890

ANNO 1841.

NUMMER van Opgaaf. *Stuger geboren Bruining Wed. J.L.*

NAMES.	GESLACHT.		OUDERDOM.	DATUM VAN AANGIFTE DER MUTATIES.			MUTATIEN EN DERVELVER RESULTATEN VAN VERMEERDERING VERMINDERING				Aanmerkingen.
	Mannljk.	Vrouwljk.		Jaar.	Dag.	Maand.	Geloorte.	VERMEERDERING		Verkoop of anderen titel.	
								Overlijden	Verkoop of anderen titel.		
<i>Cornelia</i>		X		1841	1	April					<i>Van de overleden onder naam van: Montina Hogberg, v. C.</i>
<i>Unico August</i>		X		1841	7	April					<i>Unico August Hogberg, v. C.</i>
<i>Edward Charles</i>		X		1841	7	April					<i>Edward Charles Hogberg, v. C.</i>
<i>Charlotte Christina</i>		X		1841	7	April					<i>Charlotte Christina Hogberg, v. C.</i>
<i>Anna Dorcas</i>		X		1841	7	April					<i>Anna Dorcas Hogberg, v. C.</i>
<i>Elizabeth Henriette</i>		X		1841	7	April					<i>Elizabeth Henriette Hogberg, v. C.</i>
<i>Gertruida Esther</i>		X		1841	7	April					<i>Gertruida Esther Hogberg, v. C.</i>
<i>Annette Josephine</i>		X		1841	7	April					<i>Annette Josephine Hogberg, v. C.</i>
<i>Theodoris Hugo</i>		X		1841	7	April					<i>Theodoris Hugo Hogberg, v. C.</i>
<i>Pemiere</i>		X		"	7	July					
<i>Sina</i>		/		"	"	"					
<i>Leid</i>		X		1845	20	September					
<i>Mimie</i>		/		1845	20	April					
<i>Flink</i>		X		1845	10	October					
<i>Arantuur</i>		X		1846	21	February					
<i>Adriaan</i>		/		"	"	"					
<i>Marius</i>		/		"	"	"					
<i>Pinto</i>		X		"	5	August					

Note: Entry in the slave register series 2 for the private slave owner widow L.J. Stuger née Bruining. On top, the name of the slave owner and the page number ("folio number") are mentioned, and the year this record started. Underneath from left to right, there are columns for names of enslaved, sex, date of mutation, the most common types of mutations (birth, purchase, death, sale) and remarks. People sold by the widow Stuger were struck off this entry and transferred to the entry of the new owner. Because information on enrollment and deregistration had to be entered on the same line in the register, the entries could easily become cluttered, as this example shows. For this reason, the registers were renewed every 4 to 10 years. This specific type of register was in use between 1838 and 1848 (NAS inventory number 41, folio 4890).

Roughly 75% of all enslaved were part of a plantation. In Suriname this meant that the workforce had to stay together. Plantation owners were allowed to buy and sell plantations, including the enslaved workforce, but they were not allowed to sell off enslaved individuals from their plantations without the explicit permission of the government. Because of this special legal status, a strict distinction was maintained between plantations and private slave owners, who were registered in separate books.

The other 25% of the enslaved population were privately owned. They could be bought and sold freely within the colony, as long as mothers and their children were not sold separately. Privately owned enslaved persons sold to plantations changed legal status. Until the start of series 4 in 1851, people owned by the government were not registered in the slave registers. This was deemed unnecessary because the government itself did not pay head tax and did not consider itself at risk for involvement in illegal slave trade. In the 1850s, the percentage of government owned people was roughly 1.3% of all enslaved. They mostly worked on the government plantation Catharina Sophia. Others worked for the army or as servants for the government.

All series taken together, the slave registers consisted of 56 volumes, split into 24 books for plantations and 32 books for private slave owners, see Table 1. Although privately owned enslaved persons were a minority, the larger number of books for private slave owners was necessary, because each individual slave owner had a page (folio) in the register, even if she or he owned only one person. Furthermore, names of private slave owners changed frequently over time, while plantations mostly stayed in existence and only a few new plantations were added between 1830 and 1863. Of the 56 books, 15 volumes did not survive and some books are very damaged, mainly from the older series. The newer series are by far the best preserved. Series 4 is almost complete, save for one supplement volume of private slave owners. Conversely, only one damaged volume survives of the plantation books of series 1.

Recording enslaved persons was done by plantation managers and slave owners. Slave owners in Paramaribo were given two weeks and managers of plantations a month to register any changes. However, the government was rather lenient towards them: even in the 1850s some plantations registered changes only twice a year. Because slave owners were required to record only living enslaved persons, there is a strong under-registration of infants who died before registration.

The documentation of enslaved persons of one year of age and older seems to have been rather complete. A survey of the slave registers shows that, beginning in 1859, an administrative audit was conducted in anticipation of the abolition of slavery. In the process, only 12 persons were enrolled who had not yet been registered before. There was even a tendency towards over-registration: 324 people were deleted from the registers who had died before 1859, around 0.9% of the total number of enslaved people at the time. In 1862, slave owners and plantations had to provide lists of all persons they owned to claim compensation from the government when slavery ended on 1 July 1863. When these lists, called *borderellen van aangifte*, were checked by special committees in the spring of 1863, only 31 enslaved persons were found who had not been listed in the slave registers. However, the committees found more than 830 persons missing who had escaped slavery, sometimes years or even decades earlier, but who were kept in the registers by their former owners, presumably to keep a claim on these people in case they were caught (*Algemene Rekenkamer, 1862*). This group consisted of 2.5% of the registered plantation workers and 1.5% of the registered privately owned enslaved.

Table 1 *Number of slave registers of plantations and private slave owners*

	Private owners			Plantations		
	Original number	Surviving books	Missing books	Existing books	Surviving books	Missing books
Series 1 1830–1838	9	6	3	5	1	4
Series 2 1838–1848	9	7	2	5	3	2
Series 3 1848–1851	6	5	1	6	4	2
Series 4 1851–1863	8	7	1	8	8	0
Total	32	25	7	24	16	8

See also *van Galen & Hassankhan, 2018, p. 510*.

Table 2 *Overview of information in the slave and emancipation registers*

	Slavery registers				Borderellen	Register of names
	1830–1838	1838–1848	1848–1851	1851–1863	1862	1863
<i>Owner info</i>						
Name ¹	X	X	X	X	X	X
Mutations	X	X	X	X		
<i>Personal information</i>						
First name before emancipation	X	X	X	X	X	X
First name after emancipation						X
Surname						X
Sex	X	X	X	X		
Name mother	Only new-borns	Only new-borns	X	X		X ²
Age	Only at first entrance				X	X ³
Year of birth			X	X		X ³
Profession					X	
Religion					X	
Remarks ⁴	X	X	X	X	X	X
<i>Event information</i>						
Date of registration birth	X	X	X	X		
Birth date			From 1850	X		
Date of registration death	X	X	X	X		
Date of death			From 1850	X		Sometimes added
Date of transfer	X	X	X	X		
Transfer from	X	X	X	X		
Transfer to	X	X	X	X		
Date of Manumission	X	X	X	X		
Surname after manumission	X	X	X	X		
Pawns or legal claims	X ⁵	Until 1840				

1 The name field of the owner is not differentiated and can contain information on civil status, inheritance, legal guardianship, legal representation, maiden names, multiple owners, and organizations. Names of plantations also include the location of the plantation.

2 Information on family relations was sometimes added.

3 Could be either age or year of birth.

4 Mostly used to add extra information, but additional information for which no columns existed was sometimes added, mainly on ownership or the health condition of the enslaved.

5 This contains information on people who were pawned or taken into execution for a legal claim on their owners.

Registered information differed between series, as shown in Table 2. In series 1, the 1830–1838 series, there seems to be no logical order in which slave owners and plantations were listed. Beginning with series 2 plantations and slave owners were entered in alphabetical order at the start of the series, listing information on their "property". New slave owners were added as supplements. Slave-owners were obliged to report ownership information and mutations due to births, deaths, sales, and manumissions. Initially, the registers contained little personal information, as only the name and sex of the enslaved persons were recorded, as well as the name of the mother of each new-born child. Age was only mentioned at the start of series 1 in 1830. In series 3 and 4, from 1848 onwards, the year of birth and the name of the mother was registered for all enslaved persons. If multiple persons on a plantation had the same name, a number or sometimes a personal characteristic, such as height, stature, occupation, or skin colour was added to their names to distinguish between individuals. This information was omitted when one of the individuals died or was sold to another owner.

The slave registers differ from other population registers, because of the legal status of enslaved people. Enslaved persons were first and foremost seen as commodities, meaning that they were subjected to different laws than free persons. In Suriname, they were not allowed to marry and fathers had no legal rights over their children. Paternity and other family relationships were never registered and fathers could be separated from their children if the owner wished to sell them. However, according to Surinamese law children could not be traded without permission from the government separately as long as their mother lived, not even when these children had become adults. This meant that the names of the mothers had to be registered. At the start of the slave registers, this was only done for new-born children, but from series 3 onwards registering the name of the mother was compulsory for all enslaved persons. For a more in-depth discussion of the content of the slave registers, see van Galen and Hassankhan (2018).

In 1862–1863, when the abolition of slavery was imminent in Suriname, two new types of registration were generated. In the fall of 1862, slave owners or their representatives had to hand in lists of the people they owned, in order to claim a compensation of 300 guilders per enslaved person from the Dutch government. These lists, called *Borderellen*, were structured exactly in the same way as the slave registers, but added information on religion and occupation of each enslaved person. Furthermore, in May and June 1863 a register of names was created for each district in which the emancipated former enslaved were registered with their new family name, first names, year of birth, name they had before 1863, residence and sometimes information on family relations. The information in these two sources was combined by Lamur (2004) in one single emancipation dataset. For a more in-depth discussion see Lamur (2004, pp. XVII–XLIX). The entries of roughly 34,000 emancipated individuals are linked to series 4 of the slave registers with the permission of the author.

4 DATA ENTRY OF THE SLAVE REGISTERS

An open science project was started on the online citizen science platform Vele Handen to transcribe the slave registers.¹ After registration, people got access to the online platform on which they could immediately start transcribing the assigned scan. The platform also provided general information about the project, a manual with an explanation of the data entry and transcription rules, documents to support the transcription process (such as a list with names of plantations), a forum, and the state of affairs of the project's progress. The forum, where people could interact on content-related questions, increased people's motivation and commitment to the project. It was important to constitute a community that felt responsible for providing the best transcription possible and successfully finishing the project. This community-feeling was further stimulated by regular project meetings on several locations (see van Galen (2019) for an extended discussion).

The slave registers were a suitable source for transcription on a citizen science platform, because information is already largely structured in the original documents. The structure of the source was used for the form in which participants transcribed the text (see Figure 4). Because of differences in the structure and content of the forms used by the clerks over time (see the preceding section), slightly different forms were used for each series.

1 https://velehanden.nl/projecten/bekijk/details/project/run_slavenregisters

Figure 4 Transcription interface on the Vele Handen platform



Note: The online platform Vele Handen was used to transcribe the scans of the slave registers. Volunteers were shown a scan with the fields to be filled in underneath. This was done in the same order as the original register. The register shown here is the register as it was in use in series 3 and 4 (1848–1863). To solve the lack of space mentioned in Figure 3, the information for each enslaved person was spread over two pages in these records. At the top is the page number (“folio number”) and the name of the slave owners, in this case minors J.C. and E.M. Lobato. Below that, from left to right, columns for name, sex, year of birth, mother's name, date of declaration of mutation, three columns on the addition of the enslaved (birth, purchase and remarks), three columns on the departure of the enslaved (death, sale and remarks) and a column for general remarks (NAS Slave Registers inventory number 14, folio 945).

The quality of data entry was ensured by strategies that optimized correct reading of the sources and minimized interpretation differences between volunteers. Minimizing interpretation differences among different participants depends on clear agreement. These were provided by a detailed manual, accessible in different places, e.g., via a welcome-mail, on the platform, and as a link in the transcription-form. People were urged to transcribe the exact text from the sources, for example, transcribing abbreviations, capitals and dots similar to the sources. An important exception was information that referred to multiple persons entered on the same line by means of a bracket: this information was entered for each person separately.

Differences of interpretation were minimized by having a transcription-form that corresponded to the lay-out of the source as much as possible. Source columns were split into additional separate fields only in rare cases of complex information. For example, as shown in Figure 4, all information on the name of a plantation and its owners was found in the heading of the source, for which only one field in the transcription form was reserved. During the data cleaning process (see next paragraph), this information was split into separate fields, such as initials or first names of the ‘owners’, their last names, and the plantation names. Although this method made the cleaning process more labour-intensive in some ways, it avoided a lot of discussion and differences of interpretation among volunteers, such as which initials belonged to which last name, which improved both the atmosphere among volunteers and the quality of the data entered.

Support to encourage correct reading of the sources was offered in different ways. Teaching aids for reading 19th century script were provided via the platform, such as auxiliary documents with abbreviations and words specific for the Surinamese Slave Registers and the list of plantation names mentioned earlier. Furthermore, questions could be asked on the online forum, which would be answered within a day by a fellow-volunteer or one of the project leaders. The project leaders checked the information on the forum at least once a day to ensure correct information was being spread. If people felt uncomfortable using the forum, they could always contact the project leaders by e-mail. The forum was also used by project leaders to point out frequently made mistakes in order to prevent them in the future, and to provide positive feedback to motivate people and increase commitment to providing high quality data entry.

Most important was the system of working with two different transcribers for each scan, whose transcriptions were checked by a third volunteer. This procedure ensured a correct transcription and uniformity in interpretation. These quality controllers were a small group of people selected by the project leaders based on the high quality of their work. Thanks to this system, new participants could familiarize themselves with the sources and start entering data immediately without this affecting the quality of final data entry. Because of privacy reasons, the volunteers who had entered the data were anonymized for controllers other than project leaders. The controllers became experts who functioned as stand-ins for the project leaders; they identified common mistakes which were communicated to all volunteers. This system of two separate 'data entry volunteers' and one controller per scan has been successful in data quality assurance for many other projects at the Vele Handen platform (De Moor, Rijpma, & Prats López, 2019).

5 RECORD CLEANING

5.1 SLAVE REGISTERS

The transcribed data needed to be cleaned to make the slave registers comparable and matchable. Each row in the source was kept as a record in the database. Each row referred to an enslaved person connected to a specific "owner" in one of the series. This meant that persons could return in multiple rows, if they changed "owner", or when all enslaved persons were transported to a new series. During the matching process, these records were combined into person reconstructions (see Section 6).

Before we started cleaning, all entries were assigned unique identifiers. Cleaning was strictly separated from the data entry process, so that volunteers copied all existing irregularities from the source, but did not accidentally add new variations to the database. Whenever volunteers noted irregularities in the source, they could post about it on the forum or leave a comment via a button during data entry. We used this information to write scripts that automatically detected problematic records. In some cases, this meant that we had to realign data, as the clerk — or transcriber — clearly wrote down information in the wrong column or row. For example, when dates of entry occurred after dates of exit or names occurred in date fields. Text referring to information mentioned earlier, such as ditto marks and words like "etcetera" or "as above" were replaced by the information referred to. After realigning the data, we moved on to cleaning specific variables.²

5.1.1 RECONSTRUCTING SERIES 4

Our first concern was to make series 4 as complete as possible. Series 4 was by far the most complete with only the first and last pages or entries at the bottom or top of the pages missing for the plantations, and one book missing for the private owners. As the order of the registrations in the books was consistent between series 3 and series 4, and the emancipation registers, we could deduce which plantations and concomitant enslaved people were missing. We added the names of the missing enslaved persons to series 4, and indicated whether they were alive at the start and end of the observation period. The missing book with private owners was not reconstructed, as it contained registrations on newly established slave owners from December 1859 until abolition in July 1863 and this information is readily available in the Borderellen and emancipation registers.

5.1.2 NAMES OF ENSLAVED AND THEIR MOTHERS

There was considerable variation in the spelling of names between the four series of the slave register. We left the core names intact to stay close to the original source, but redundant punctuation marks were removed, such as full stops after names. To make matching easier, we split person descriptions from person names by writing out all abbreviations and moving nicknames referring to attributes (like age, colour, occupation, plantations, or stature), numerals, and unidentifiable abbreviations to a separate variable. If the record indicated that a person was known by multiple first names, we opted to keep all name variations in the same field and separate them with the phrase *of*, the Dutch word for 'or'.

² For more details, see <https://www.ru.nl/hdsc/online-sources/suriname-slave-registers/>.

5.1.3 NAMES OF OWNERS

Plantation names were standardized by removing information on produce and location. Then, the names of plantations were standardized using almanacs that list the names of all ~370 plantations that existed between 1830 and 1863 ([Maatschappij tot Nut van 't Algemeen, 1830, 1846](#); [Ministerie van Koloniën, 1856](#)). These almanacs were also used to create an additional dataset of Surinamese plantations, which is made available with the slave registers dataset ([Rosenbaum-Feldbrügge et al., 2023](#)). Cleaning the names of private owners required manual review, as the order of first and last names was not fixed and the name field can contain information on civil status, inheritance, legal guardianship, legal representation, maiden names, multiple owners, and organizations.

Private owners consisted of both natural and juridical persons. Organization names were written out in their entirety, and persons' names required strict standardization. We placed the family name of the first-mentioned owner at the start of the name and removed repeated instances if multiple owners had the same family name.³ Redundant punctuation marks were removed. Information on civil status, inheritance, or maiden names was standardized by removing the different variations of *weduwe* 'widow', *boedel* 'inventory', and *geboren* 'née'. Legal guardians were flagged. We separated owners from legal representatives. Information on the owners was put in the normal text fields, while information on representatives was flagged with the phrase ", door ... qq" to extract them. Most commonly these were husbands who represented their spouse. However, it also happened that a parent acted as the legal representative of a child's trust fund, or someone acted as the legal representative of an inheritance.

5.1.4 SEX

Information on sex from the slave registers was recoded into three possible values: female, male, or unknown. To reduce unknown and miscoded data, we checked whether each name was uniformly coded as female, male, or unknown and flagged all 571 names that were not. We manually controlled these entries and found that when <25% had the alternative sex, the sex of the enslaved person was almost certainly wrongly coded. Therefore, we decided to automatically recode the 440 names where <25% had the alternative sex, except for four names that could be used for both sexes, and used the same procedure to resolve the unknowns.

5.1.5 IN AND OUT EVENTS

For every record, we made separate fields for "in event" and "out event" to clarify why and when observation of the enslaved person started and ended. In and out events were identified using string searches on the text fields that accompanied mutations, such as *vrijheid* 'freedom', *geb* 'born abbr.', or *verk* 'sold abbr.'. We were able to retrieve nearly 100% of all in and out events and could date them accurately. In principle there were three main start events (birth, start of the series, and transfer) and three main end events (death, end of the series, transfer) that comprised more than 96% of all mutations. The remaining 4% mutations were corrections of errors in the register or removal due to disease, ill health, escape, or manumission of the enslaved before 1 July 1863 (the official abolition of slavery in Suriname marking the end of series 4).

5.1.6 DATES

Dates related to the in and out events were standardized to the yyyy-mm-dd format. These dates were directly mentioned in the source itself, except for dates marking the beginning and end of a series and most dates of birth in series 1 and 2. In those two series, dates were mentioned for new-borns only; for others, only the age at registration was available, so date of birth was estimated by subtracting age from the registration date.

3 In theory, each folio contains only one owner, or one set of owners, but in practice there are ample exceptions to this rule. In those cases, clerks wrote down extra information on the folio, struck out owner names, or added new ones. The clerks sometimes chose this solution when the enslaved were not sold to a third party, but inherited by a widow, inherited by offspring, marked as inventory for inheritance, or sold to a secondary owner already on the folio. We chose not to split these records, as the change in ownership was not dated. Rather, we sort ownership by the family name of the first owner, so that we follow the logic of the clerks.

We ensured that there are no missing start and end dates and that there are no date inconsistencies (such as individuals dying before they are born or individuals leaving before they enter), which were mainly caused by incorrect date transcriptions.

5.1.7 REMARKS

Remarks in the source were stored in a separate field, which received very little cleaning, because of a large variety of types of remarks and formulations. Only references to information mentioned with foregoing persons in the source, for example by ditto marks, was used to add information, so each entry would be understandable in itself. The remarks were used to retrieve information on in and out events (see Section 4.1.5), and might be standardized further in later versions of the database. In the meantime, the remarks (written in Dutch) are understandable in their current form.

5.2 EMANCIPATION REGISTER

The emancipation register dataset was constructed by Lamur (2004) from the *Borderellen* and the register of names that was generated for each Surinamese district after emancipation (see [Algemene Rekenkamer, 1862](#)). Accordingly, some basic cleaning had already been applied by the author, and the plantation names had already been standardized. Nevertheless, it was necessary to correct obvious typos and inconsistencies in the Lamur dataset and split or standardize certain variables. The name of the enslaved before 1863, for instance, was split into first names, nicknames based on attributes, and baptismal name, if applicable. The private owners' name was split into the last name and the remaining information. Similarly, information on residential locations, enslaved names, occupations, and other remarks were split into multiple variables.

6 MATCHING

6.1 MATCHING INDIVIDUALS TO RECONSTRUCT LIFE COURSES

The main feature of the reconstituted life courses when compared to the Suriname: Slavenregisters Dataset 1830–1863 already published with the Dutch National Archives is that we matched individuals within and between the four series of the slave register as well as with the emancipation register. Accordingly, we reconstructed individual life courses of the enslaved population of Suriname which enables researchers to conduct longitudinal life course research.

Matching within the series was necessary to follow enslaved individuals who were transferred from one owner to another, because they were sold, given away, or inherited. Matching between the series was necessary to follow enslaved individuals that lived with their owners when a new series was created, that is, 1838, 1848, and 1851. To give an example, Philippina (Id_person: 24082) was born on plantation Anna Catharina in 1849 when series 3 was still in place. In 1851, series 4 started which required matching Philippina *between* series 3 and 4. In 1861, Philippina was transferred from plantation Anna Catharina to plantation Kroonenbrug which required matching her *within* series 4. When slavery was abolished in July 1863, 14-year-old Philippina was still alive and successfully matched to the emancipation register.

6.2 MATCHING THE RECORDS IN THE SLAVE AND EMANCIPATION REGISTERS

To account for differences in spelling (for instance Philippina, Philipina, Philippine), we matched names of enslaved, their mothers, and private owners with a maximum Levenshtein distance dependent on the length of the specific name. Names of plantations were matched without Levenshtein distances, as they were already standardized. Our matching algorithm ignores internal blanks and upper-case characters. In addition, certain letters and letter combinations were replaced, such as kw => qu and ph => f), as they were used interchangeably in the register.

The matching process proceeds in several steps. In the first step, we select the relevant entries based on the in and out events. In the second step, we retrieve candidate matches based on the Levenshtein distances between enslaved names using the "property to person" algorithm ([Mourits & Rosenbaum-Feldbrügge, 2023](#)). Third, we filter out obviously false matches, based on certain rules. Most

importantly, we discarded matches between females and males as well as matches with different birth years. When matching different series, we also discarded matches of enslaved persons who belonged to different owners. In the fourth step, we scored matches on their plausibility using characteristics such as mother's name, nicknames, year of birth, and names of the enslaved in the preceding and following entry (the order of names was often kept when new registers were made or several enslaved were transferred at the same time). For instance, correct mother names were awarded 2.5 points, correct year of birth 2 points, and name of enslaved in preceding or proceeding entry 1 point.⁴ Finally, we selected the match with the highest score per entry, and discarded ties or matches with low scores to prevent false matches.⁵

The matching approach differed slightly for matching between series and matching within series. For matching between series, we included records that were present at the end of one series or the start of the following series, and only matched records with the same owner. For matching within series, in contrast, we only selected entries whose out event or start event was a transfer, and matched *only* on the name of the enslaved (because the enslaved changed owners). In addition, we filtered candidate matches also based on the date of transfer. To deal with wrongfully transcribed years, we also allowed matches with an identical month and day of birth instead of a matching year of birth, and matches with an identical month and day of transfer instead of a matching year of transfer.

The end of series 4 and the emancipation register were matched according to the same procedure as applied for the between matches. First, we included records that were present at the end of series 4 and all records in the emancipation registers. Second, we matched records based on the first name of the enslaved and the name of the owner. Third, we filtered out individuals with an identical or unknown sex and an identical or unknown year of birth. Since sex is not available in the emancipation register, we inferred the sex of the individuals based on their first names before starting the matching procedure. To achieve this, we used information on sex and name combinations derived from the slave registers. We did not use mother's name in our matching procedure, because it was usually not available in the emancipation register. Finally, we picked the highest-scoring entry and dropped ties. In a second step, that we only applied to the remaining unmatched cases, we also allowed matches between privately-owned and plantation-owned individuals, as information on both plantations and the plantation owners is available in the emancipation register dataset.⁶

6.3 CHECKING RECOND LINKAGE QUALITY

We checked the quality of the matching procedure in three steps. First, we checked the retrieval by selecting a limited number of plantations and private owners to check whether unmatched records could be matched by hand. Second, we checked the precision by verifying whether the established matches made sense. Finally, we checked the reconstructed life courses for inconsistencies to see if parts of our process needed reiteration. These tests highlighted several challenges in our historical data that needed to be dealt with.

6.3.1 PROBLEMS AND SOLUTIONS

Since the slave registers consist of four series that registered enslaved persons in slightly different ways, this caused some difficulties when matching the records of an enslaved person from one series to those of him or her in the following series. Another challenge was caused by the fact that not all plantations and private owners are present in all four series, as not all registers survived. Obviously, all the records of a plantation that was only present in one of the two series would always turn out as missing matches. For instance, when plantation Arendrust is not present in series 1 while it is part of series 2, all records of plantation Arendrust will not have a match when matches are made between series 1 and 2. Therefore, we took a sample of 44 plantations to check for missing matches between two series. For each set of matched series different plantations were selected to minimize biases. The matching rate in the samples ranged from 81% for individuals who survived series 1 to 99% who were transferred from series 3 to series 4. This indicates that our matching algorithm works better for the later slave registers, which is logical, seeing that the latter series contain more information on the year of birth and mother of the enslaved.

4 For more details, see <https://www.ru.nl/hdsc/online-sources/suriname-slave-registers/>.

5 The entire script is available at <https://github.com/HDSC-Nijmegen/Slavenregisters>.

6 For more details, see <https://www.ru.nl/hdsc/online-sources/suriname-slave-registers/>.

Although the algorithm automatically matched the vast majority of records correctly, some recurring problems emerged from unstandardized names. Records of enslaved on plantations match better than records of privately owned slaves, as the names of plantations were all standardized. For the private owners, variations in the spelling of the last names sometimes led to overlooked matches. Similarly, the lack of standardization for the spelling of both names of enslaved people and mother's names was initially a major cause of missing matches. Ignoring whitespaces and upper cases solved most of this problem. But all mismatches caused by spelling variations could not be solved, because most names cannot be standardized without manually checking the source. Allowing too many letters to differ would, especially for shorter names, result in more incorrect matches being added than correct matches being gained. For example, this could result in Damon and Simon being matched whilst they are clearly two different names. Therefore, missed matches can also be seen as an indication of transcription quality.

Another recurring obstacle was a difference in birth year between two records. Often there was only a one-year discrepancy in the birth year when the series reported the age at registration rather than the birth date. This issue could be solved relatively easily by allowing a one-year difference while filtering our candidate matches. However, problems were harder to solve when years or dates contained transcription errors. Levenshtein distances do not apply to dates, so getting a single digit wrong results in a missed match if date filters are applied stringently. On the other hand, dates cannot be ignored, as this would steeply increase the number of false matches. To retrieve records with transcription errors in the dates, we applied extra filtering rules, such as having matching names of the enslaved in the preceding and following entry or having the same month and day of birth.

More troublesome was missing information. When a birth year was missing in one of the series, the candidate match received a lower matching score. Depending on the scores for other elements this could result in a missing match. This problem could also surface with other matching elements, for instance the mother's name was present in series one and not in series two. This problem occurred most often in the earlier series as these contained less information. This issue was partly solved by allowing missing mother names to also match. Nevertheless, the quality of the results from our matching algorithm is highly dependent on the available information, which increases with each series.

It was much harder to adapt the algorithm to cases where multiple factors prevented records from being linked. These complex cases required close comparison of the two records and consultation of the primary documents, as they indicate the limitations of the matching program. Fortunately, the complex cases mostly occurred on a small number of plantations where the registration was not done correctly. Again, the quality of the matching result is an indication of the quality of registration and transcription.

6.3.2 MATCHING WITHIN THE SERIES

Table 3 shows that the share of fully observed lives in slavery within a series, which are defined as life courses that could be completely followed within the relevant series, increases over time ranging from 57.4% in series 1 to 93.4% in series 3. That the first two series have a significantly lower share of fully observed life courses is a consequence of the higher number of missing registers (see Table 1) and lower rates of matching given fewer types of information available. With 93.4% and 92.7%, series 3 and 4 have a high share of complete life courses indicating high internal consistency.

Table 3 *Number and percentage of complete life courses within series*

Series	Total life courses	Complete life courses in %
Series 1: 1830–1838	21,939	57.4
Series 2: 1838–1848	33,319	70.1
Series 3: 1848–1851	33,626	93.4
Series 4: 1851–1863	55,644	92.7

6.3.3 MATCHING BETWEEN THE SERIES

Table 4 shows the share of individuals that are successfully matched between the end of the preceding and the beginning of the succeeding series. Matching success between the first and the second series is comparatively low given the large number of missing books and the scarce personal information available in these series. Matching success is highest between series 3 and series 4 with nearly 96% of enslaved individuals present at the end of series 3 being matched to individuals present at the beginning of series 4.

Table 4 *Number and percentage of persons matched between series*

Series	Number of persons matched	Persons present at end of preceding series		Persons present at beginning of preceding series	
		Number	% Matched	Number	% Matched
Between series 1 and series 2	4,889	13,699	35.7	21,263	23.0
Between series 2 and series 3	11,924	17,800	67.0	29,449	40.5
Between series 3 and series 4	27,741	28,938	95.9	39,554	70.2

6.3.4 MATCHING BETWEEN SERIES 4 AND EMANCIPATION REGISTER

Compared to the internal matching of the slave registers, the matching between the end of series 4 and the emancipation register has a high success rate. Table 5 below shows that more than 90% of the enslaved individuals present at the end of series 4 and nearly 88% of the individuals present in the emancipation register are matched.

Table 5 *Number and percentage of matched life courses between series 4 and emancipation register*

Series	Persons matched	Persons present at end of series 4		Persons present in emancipation register	
	Number	Number	% Matched	Number	% Matched
Between series 4 and emancipation register	30,133	33,432	90.1	34,430	87.5

We checked all matches between series 4 and the emancipation registers manually and found not more than 73 incorrect (0.2%) and 65 unclear matches (0.2%), indicating that the matching quality between series 4 and the emancipation register is extraordinarily high.

7 THE OUTPUT

The database Suriname Slave and Emancipation Registers Dataset is stored in CSV-format in UTF-8 code according to open standards, with commas (,) as separators and quotation marks (" ") to signal text fields. The database and the detailed description can be downloaded from the IISH Dataverse ([Rosenbaum-Feldbrügge et al., 2023](#)). The database contains one table in which each row in the data represents a unique entry in the slave registers for one individual. Individuals have multiple entries in the registers if they were sold to another owner or appeared in several series. The life course of an enslaved person between 1830 and 1863 can be reconstructed by linking all entries using the `Id_person` variable. An additional database Suriname Plantation Dataset with information on plantations in Suriname is included for researchers who want to do research on specific types of plantations, such as selecting by produce or location ([Rosenbaum-Feldbrügge et al., 2023](#)).

The dataset provides basic demographic information on date of birth, date of death, sex of the enslaved, name of the enslaved, mother's name, and the name of the owner. In addition, the start and end date of each entry as well as the reason for the start and the end of each entry is recorded. A full list of variables, their short descriptions, and their variable type can be found in the Appendix.

8 POSSIBILITIES AND LIMITATIONS OF THE SURINAME SLAVE AND EMANCIPATION REGISTERS DATASET

Until now, all research on the social and demographic history of slavery in Suriname had to rely on case studies. Often, these were relatively large plantations or the government plantation Catharina Sophia, but these might not be representative for the enslaved population as a whole (Everaert, 1999; Lamur, 1987; van den Boogaart & Emmer, 1977; van Stipriaan, 1993). The current dataset enables researchers to study not only all plantations, but also often overlooked groups of enslaved persons, like household staff and craftsmen working in the cities.

The completion of the database will expand the scope of slavery research. This opens the opportunity for demographic research to compare fertility and mortality across different types of enslaved and different types or locations of plantations. The social history of slavery can also be enriched, as we have more insight on who was living in slavery, who managed to escape or who was manumitted. To what extent were formerly enslaved able to manumit their friends and family, for example?

Nevertheless, there are some limitations to the current dataset. First, the linkage of the various entries per enslaved person is good but not perfect. Due to lost slave registers, misspellings and a conservative matching approach, optimal linking has not yet been achieved. Because the linking strategy was probability-based, a threshold was used to match entries based on a linking score. It is impossible to systematically assess whether all individual links are correct, but smaller case studies indicate the share of incorrect links is quite low (around 1%). As the number of persons in the database is large enough, it is still possible to provide reliable outcomes. Series 1 and 2 are much less complete and contain less individual information, which strongly decreases the share of correct matches. The missing registers are random. In most cases, the absence of this material does not bias research conducted with it. The only exception is research on plantations in the 1830s. The only remaining plantation register in Series 1 contains mostly plantations from the western districts of Nickerie and Coronie. These were new plantations that were only eight to twenty years old in 1830. As during the construction phase mostly men were purchased, the gender ratio is not typical for Surinamese plantations.

Other limitations are the lack of information on ethnicity or occupation in the slave registers. Occupation is only available for enslaved persons living in 1862–1863. Furthermore, the slave registers are based on legal ownership. This meant that enslaved persons were connected to their legal owner, a plantation or a private person, but they could be rented out to others. Particularly for the privately owned enslaved, it is unclear whether the address of the owner always corresponded with the place where the enslaved lived. This makes it difficult to study the external conditions in which these people lived. Were they enslaved in an urban environment or engaged in plantation labour, and in what kind of area in the country?

For the persons enslaved on plantations we can be more certain of the places where they lived and performed their day jobs. They mostly stayed on the same plantations and when they were transferred to other plantations it was often done for the whole group at once and registered in the slave registers. For this group it is therefore possible to examine the external conditions of their enslaved lives.

To some extent missing information can be complemented by enriching the information in the slave registers by linking to other sources. For example, in the 1850s the Surinamese almanacs offer information on the whereabouts of groups of plantation workers who were rented out to other plantations. In Paramaribo, the only city in the colony of Suriname, we can enrich the information in the slave registers by linking to other sources, for instance the Paramaribo ward registers. These annual registers recorded all inhabitants of the capital per household, also counting the number of enslaved people domiciled at an address, including information on sex, skin color and whether they were child or adult. In combination with the names of the free persons in the household (usually the head of the household) as potential

owners, we can match around 43% of the enslaved persons from the slave registers to the ward registers. In this way, we can establish with a reasonable degree of certainty where urban enslaved people lived.⁷

9 CONCLUSION

The publication of the database of the Surinamese slave and emancipation registers is a step in the ongoing process of forming the Historical Database of Suriname and the Caribbean (HDSC). Emphatically, it remains a work in progress. In the coming years we will continue to improve the existing database. Linkage is already exceeding 90% for the slave registers from 1848 onwards, but our goal is to increase it even further by addressing specific issues. First of all, we will improve the cleaning process by dealing with the missing dates and inconsistent information described in the cleaning section. Second, we will improve our matching algorithm to further increase linkage success. In particular, we will allow for a larger difference between birth years to address age heaping in series 1. Finally, we will improve our matching algorithm with manual checks and corrections on the underlying transcribed data to minimize incorrect and maximize correct matches.

We also want to further improve functionality for users by improving the linkage to the mothers of enslaved people and by standardizing information, including information on private slave owners and occupations. In the long-term, we see the HDSC leading to a Caribbean Demographic Database, comprising both Dutch and non-Dutch territories in the Caribbean. This database will function as a data infrastructure from which tailor-made databases can be formed, adapted to the specific research questions of researchers from history, sociology and other disciplines.

ACKNOWLEDGEMENTS

The development of the database of the Surinamese slave and emancipation registers could not have been possible without the support of more than 600 volunteers and a number of specialists like Dr. Maurits Hassankhan, emeritus professor Humphrey Lamur and the other board members of the HDSC Foundation. We are very grateful for their support. Our gratitude also extends towards the National Archives of Suriname, the National Archives of the Netherlands, the Radboud University Nijmegen and the Anton de Kom University of Suriname for their practical and institutional support.

In addition to the almost 400 private donors who supported our crowd funding campaign, we received funding from PDI-SSH, the Gerda Henkel Stiftung, Prins Bernhard Cultuurfonds, Stichting Democratie en Media, CLARIAH and the Radboud Institute for Culture and History (RICH).

REFERENCES

- Algemene Rekenkamer (1862). 2.02.09.08 *Inventaris van het archief van de Algemene Rekenkamer 1814–1919: Comptabel beheer. 223–248 Stukken tot opheffing der slavernij in West-Indië 1863–1868* [2.02.09.08 Inventory of the Court of Audit's archives 1814–1919]. Den Haag: Nationaal Archief. Retrieved from <https://www.nationaalarchief.nl/onderzoeken/archief/2.02.09.08/invnr/%40235~223-248>
- Arends, R. M., Raaijmakers, W., Rosenbaum-Feldbrügge, M., & van Galen, C. W. (2023). *Aruba: Slavernij en emancipatie, 1840–1863* [Aruba: Slavery and emancipation, 1840–1863] [Database]. Retrieved from <https://www.nationaalarchief.nl/onderzoeken/index/nt00479>

7 In a pilot we selected one district in the 1846 register and could manually match enslaved people living on 41 out of 96 addresses with varying degrees of certainty. See https://gitlab.com/thunnis1/WOCE/-/blob/main/Pilot%3A%20matching%20with%20slave%20registers/Rapport_Bevindingen_Matchen_Wijkregister_en_Slavenregister.docx

- Bone, J., Archer, M., Barraclough, D., Eggleton, P., Flight, D., Head, M., Jones, D. T., Scheib, C., & Voulvoulis, N. (2012). Public participation in soil surveys: Lessons from a pilot study in England. *Environmental Science & Technology*, 46(7), 3687–3696. doi: [10.1021/es203880p](https://doi.org/10.1021/es203880p)
- Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., & Wilderman, C. (2009). *Public participation in scientific research: Defining the field and assessing its potential for informal science education*. Washington, DC: CAISE.
- Buddingh, H. (2022). *A history of Suriname*. Leiden: Sidestone Press.
- Cohn, J. P. (2008). Citizen science: Can volunteers do real research? *BioScience*, 58(3), 192–197. doi: [10.1641/B580303](https://doi.org/10.1641/B580303)
- Crall, A., Kosmala, M., Cheng, R., Brier, J., Cavalier, D., Henderson, S., & Richardson, A. (2017). Volunteer recruitment and retention in online citizen science projects using marketing strategies: Lessons from Season Spotter. *Journal of Science Communication*, 16(1), 1–29. doi: [10.22323/2.16010201](https://doi.org/10.22323/2.16010201)
- De Moor, T., Rijpma, A., & Prats López, M. (2019). Dynamics of engagement in citizen science: Results from the “Yes, I do!”-project. *Citizen Science: Theory and Practice*, 4(1), 38. doi: [10.5334/cstp.212](https://doi.org/10.5334/cstp.212)
- Eltis, D., Halbert, M., & Misevich, P. (2008). *Voyages: The trans-Atlantic slave trade database*. Retrieved from <http://www.slavevoyages.org>
- Enslaved.org. (2018). *Enslaved. Peoples of the historical slave trade*. Retrieved from <https://enslaved.org/>
- Everaert, H. A. M. (1999). *Een zoektocht naar de aard van man-vrouw relaties onder Surinaamse slaven: De suikerplantages Fairfield, Breukelerwaard, Cannewapibo en La Jalousie in de periode voorafgaande aan de emancipatie* [A search into the nature of male-female relationships among Surinamese slaves: The sugar plantations Fairfield, Breukelerwaard, Cannewapibo and La Jalousie in the period before the emancipation]. (Doctoral dissertation). Retrieved from <https://dare.uva.nl/search?identifier=60a657f5-1cff-419c-bab0-870e5b392b03>
- Everaert, H. (2013). Mogelijkheden en moeilijkheden van de levensloopbenadering in de historische demografie van Suriname [Possibilities and problems of a life course approach in the historical demography of Suriname]. In M. S. Hassankhan, J. L. Egger & E. R. Jagdew (Eds.), *Verkenningen in de historiografie van Suriname: Van koloniale geschiedenis tot geschiedenis van het volk* (pp. 207–228). Paramaribo: Anton de Kom University.
- Franzoni, C., & Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. *Research Policy*, 43(1), 1–20. doi: [10.1016/j.respol.2013.07.005](https://doi.org/10.1016/j.respol.2013.07.005)
- Hall, C., & Draper, N. (2019). *Legacies of British Slave-ownership*. Retrieved from <https://www.ucl.ac.uk/lbs/>
- Higman, B. W. (1976). *Slave population and economy in Jamaica, 1807–1834*. Cambridge: Cambridge University Press.
- Higman, B. W. (1984). *Slave populations of the British Caribbean, 1807–1834*. Baltimore: The Johns Hopkins University Press.
- Irwin, A. (1995). *Citizen science. A study of people, expertise and sustainable development*. London: Routledge.
- Kuyper, J., & Heyse, D. (ca. 1882). *Kaart van Suriname* [Map of Suriname]. S.l.: Vereeniging voor Suriname. Retrieved from <https://hdl.handle.net/11245/3.38601>
- Lamur, H. E. (1981). Demographic performance of two slave populations in the Dutch speaking Caribbean. *Boletín de Estudios Latinoamericanos y del Caribe*, 30, 87–102. Retrieved from <https://www.jstor.org/stable/25675094>
- Lamur, H. E. (1987). Fertility differentials on three slave plantations in Suriname. *Slavery & Abolition*, 8(3), 313–335. doi: [10.1080/01440398708574941](https://doi.org/10.1080/01440398708574941)
- Lamur, H. E. (2004). *Family name and kinship of emancipated slaves in Suriname: Tracing ancestors*. Amsterdam: KIT.
- Langefeld, E., van Galen, C. W., Quanjer, B., & Paul, M. (2020). *Curaçao: Slavenregister en emancipatieregisters, 1839–1863* [Curacao: Slave register and emancipation registers, 1839–1863] [Database]. Retrieved from <https://www.nationaalarchief.nl/onderzoeken/index/nt00462>
- Lovejoy, H. (2015). *Liberated Africans*. Retrieved from www.liberatedafricans.org
- Maatschappij tot Nut van 't Algemeen. (1830). *Surinaamsche Almanak voor het jaar 1831* [Surinamese Almanac for the year 1831]. Amsterdam: C.G. Sulpke. Retrieved from https://www.dbnl.org/titels/titel.php?id=_sur001183101

- Maatschappij tot Nut van 't Algemeen. (1846). *Surinaamsche Almanak voor het jaar 1847* [Surinamese Almanac for the year 1847]. Amsterdam: C.G. Sulpke. Retrieved from https://www.dbnl.org/titels/titel.php?id=_sur001184701
- Mandemakers, K., & Dillon, L. (2004). Best practices with large databases on historical populations. *Historical Methods*, 37(1), 34–38. doi: [10.3200/HMTS.37.1.34-38](https://doi.org/10.3200/HMTS.37.1.34-38)
- Mandemakers, K. (2023). “You really got me”. *Ontwikkeling en toekomst van historische databestanden met microdata* [Development and future of historical databases with microdata] (Valedictory speech). Rotterdam: Erasmus University Rotterdam. doi: [10.25397/eur.23256467](https://doi.org/10.25397/eur.23256467)
- Margo, R. A., & Steckel, R. H. (1982). The heights of American slaves: New evidence on slave nutrition and health. *Social Science History*, 6(4), 516–538. doi: [10.2307/1170974](https://doi.org/10.2307/1170974)
- Menard, R. R. (1975). The Maryland slave population, 1658 to 1730: A demographic profile of blacks in four counties. *The William and Mary Quarterly*, 32(1), 29–54. doi: [10.2307/1922593](https://doi.org/10.2307/1922593)
- Meredith John, A. (1988). *The plantation slaves of Trinidad, 1783–1816: A mathematical and demographic enquiry*. Cambridge: Cambridge University Press.
- Midlo Hall, G., Hawthorne, W., & Mitchell, B. (2019). *Slave biographies: The Atlantic database network*. Retrieved from <https://slavebiographies.org>
- Ministerie van Koloniën. (1856). *Almanak voor de Nederlandsche West-Indische bezittingen, en de kust van Guinea. Jaargang 1856* [Almanac for the Dutch West Indian territories and the coast of Guinea: 1856]. Den Haag: Gebroeders Van Cleef. Retrieved from https://www.dbnl.org/titels/titel.php?id=_alm009185601
- Mourits, R. J., & Rosenbaum-Feldbrügge, M. (2023). *Property to person*. Retrieved from <https://github.com/RJMourits/P2P>
- Raaijmakers, W., & van Galen, C. W. (2023). *Sint Eustatius: Borderellen en emancipatieregister, 1862–1863* [Sint Eustatius: Borderellen and emancipation register] [Database]. Retrieved from <https://www.nationaalarchief.nl/onderzoeken/index/nt00476>
- Rosenbaum-Feldbrügge, M., Mourits, R. J., A.B., M., Janssen, J., Quanjer, B., van Oort, T., Kok, J., & van Galen, C. W. (2023). *Suriname slave and emancipation registers dataset version 1.1* [Dataset]. Retrieved from <https://hdl.handle.net/10622/CSPBHO>
- Rosenbaum-Feldbrügge, M., Mourits, R. J., van Oort, T., & van Galen, C. W. (2023). *Suriname burgerlijke stand: Geboorteakten Paramaribo (1828–1921)* [Suriname civil records: Birth certificates of Paramaribo] [database]. Retrieved from <https://nationaalarchief.nl/onderzoeken/alle-genealogie/genealogie-burgerlijke-stand/persons>
- Rosenbaum-Feldbrügge, M., van Galen, C. W., & Swaters, D. (2023). *Suriname plantation dataset version 1.0* [Dataset]. Retrieved from <https://hdl.handle.net/10622/VTL43W>
- Sauermann, H., & Franzoni, C. (2015). Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences*, 112(3), 679–684. doi: [10.1073/pnas.1408907112](https://doi.org/10.1073/pnas.1408907112)
- van den Boogaart, E., & Emmer, P. C. (1977). Plantation slavery in Surinam in the last decade before emancipation: The case of Catharina Sophia. In: V. Rubin & A. Tuden (Eds.), *Comparative perspectives on slavery in new world societies* (pp. 205–225). New York: NY Academy of Sciences.
- van Dusseldorf, D. B. W. M. (1966). De tweede algemene volkstelling Suriname 1950 [The second general census of Suriname 1950]. *Nieuwe West-Indische Gids*, 45(1), 38–44. Retrieved from <https://www.jstor.org/stable/41970086>
- van Galen, C. W. (2019). Creating an audience: Experiences from the Surinamese slave registers crowdsourcing project. *Historical Methods*, 52(3), 178–194. doi: [10.1080/01615440.2019.1590268](https://doi.org/10.1080/01615440.2019.1590268)
- van Galen, C. W., A.B., M., Mourits, R. J., & Rosenbaum-Feldbrügge, M. (2019). *Suriname: Slavenregisters dataset 1830–1863* [Surinam: Slave registers data set 1830–1863] [Database]. Retrieved from <https://www.nationaalarchief.nl/onderzoeken/index/nt00461>
- van Galen, C. W., & Hassankhan, M. S. (2018). A research note on the slave registers of Suriname, 1830–1865. *History of the Family*, 23(3), 503–520. doi: [10.1080/1081602X.2018.1507917](https://doi.org/10.1080/1081602X.2018.1507917)
- van Stipriaan, A. (1993). *Surinaams contrast: Roofbouw en overleven in een Caraïbische plantagekolonie 1750–1863* [Surinamese contrast. Exploitation and survival in a Caribbean plantation colony 1750–1863]. Leiden: KITLV.

APPENDIX VARIABLES IN RELEASE SLAVEREGISTER_SURINAME_V1.1

Category	Variable	Description	Process
Identifiers	Id_person	Person identifier (primary key)	constructed
	Id_source	Entry identifier	transcribed
General	Name_enslaved	The name of the research person	transcribed
	Sex	Sex of the research person	logical edit
	Age	Age of the research person as stated in the register	transcribed
	Age_Sex	Categories for age and sex as stated in the register	logical edit
	Day_birth	Day of birth	constructed
	Month_birth	Month of birth	constructed
	Year_birth	Year of birth	transcribed
	Year_birht2_ER	Second year of birth from emancipation register	transcribed
	Day_death	Day of death	constructed
	Month_death	Month of death	constructed
	Year_death	Year of death	constructed
	Name_mother	Name of the mother of the enslaved person	transcribed
	Plantation	Name of the plantation for that entry	standardized
	Name_owner	Name of the owner for that entry	transcribed
	Start entry	StartEntryDay	Start day entry
StartEntryMonth		Start month entry	constructed
StartEntryYear		Start year entry	constructed
StartEntryInfo		Reason for start entry (Dutch)	transcribed
StartEntryEventDetailed		Detailed reason for start entry	constructed
StartEntryEvent		Reason for start entry	constructed
End entry	EndEntryDay	End day entry	constructed
	EndEntryMonth	End month entry	constructed
	EndEntryYear	End year entry	constructed
	EndEntryInfo	Reason for end entry (Dutch)	transcribed
	EndEntryEventDetailed	Detailed reason for end entry	constructed
	EndEntryEvent	Reason for end entry	constructed
Emancipation register	First_name	First name after emancipation	transcribed
	Family_name	Family name after emancipation	transcribed
	Baptized_name	Baptized name	transcribed
	Family_relations	Information about family members	transcribed
	Occupation	Occupation on emancipation register	transcribed
	Remarks_ER	Further remarks on emancipation register	transcribed
Source	Inventory_number	Inventory number of the original source	transcribed
	Folio_number	Folio number of the original source	transcribed
	Serieregister	Year of the register	transcribed
	Typeregister	Type of the register	transcribed

II

Family restitutions



HISTORICAL LIFE COURSE STUDIES
VOLUME 13 (2023), published 25-04-2023

PRDH and IMPQ 1800–1849 Quebec Historical Family Reconstitution

Content, Design and Biographical Completeness

Lisa Dillon

Marilyn Amorevieta-Gentil

Alain Gagnon

Bertrand Desjardins

Département de démographie, Université de Montréal

ABSTRACT

Since 1966, the Programme de recherche en démographie historique (PRDH) has worked to create comprehensive genealogical data of the Quebec population. The PRDH longitudinal database, the Registre de la population du Québec ancien (RPQA), draws upon the French Catholic parish registers of the St. Lawrence Valley as its main source material. This family reconstitution covers the French Catholic population of Quebec up to 1799, along with deaths after 1800 of persons born before 1750. Subsequent partnerships with l'Institut Généalogique Drouin, FamilySearch and Ancestry as well as collaboration on the 2011–2017 *Infrastructure intégrée des Microdonnées historiques de la Population du Québec (1621–1965)* (IMPQ) project enabled the PRDH to continue efforts to reconstitute the French Catholic population up to 1849. Despite these advances, pushing family reconstitution forward to the mid-19 century has forced the PRDH team to reckon with the increasingly mixed and geographically mobile Quebec population of the 19th and early 20th centuries. This article describes the content and design of the RPQA database, detailing the structure of the RPQA relational database and the breadth of variables available for data management and analysis. It then describes features of the IMPQ extension of family reconstitution from 1800 to 1849, including observational protocols necessary to use these data and consideration of data completeness after 1800. At the same time, the article addresses the fundamental question, "What is my population?" as part of a broader reflection upon the target population encompassed by these data.

Keywords: Registre de la population du Québec ancien (RPQA), Programme de recherche en démographie historique (PRDH), Infrastructure intégrée des microdonnées historiques de la population du Québec (IMPQ), Institut généalogique Drouin (IGD), BALSAC, Family reconstitution, Record linkage, Historical population data, Genealogical data, Parish registers, Censuses

DOI article: <https://doi.org/10.51964/hlcs13984>

© 2023, Dillon, Amorevieta-Gentil, Gagnon, Desjardins

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

Since 1966, the Programme de Recherche en Démographie Historique (PRDH) has worked to create comprehensive genealogical data of the Quebec population. The PRDH longitudinal database, the *Registre de la Population du Québec Ancien (RPQA)*, draws upon the French Catholic parish registers of the St. Lawrence Valley as its main source material. The PRDH has also incorporated into the RPQA complete-count 17th- and 18th-century census microdata as well as supplementary information drawn from notarial and other records (Dillon et al., 2018). By the early 2000s, the PRDH had succeeded in reconstituting the French Catholic population of Quebec up to 1799, along with deaths after 1800 of persons born before 1750 (Dillon et al., 2018). Subsequent partnerships with l'Institut Généalogique Drouin (IGD), BALSAC and the Centre Interuniversitaire d'Études Québécoises (CIEQ) on the 2013–2017 *Integrated infrastructure of the historical microdata of the Quebec population (1621–1965)* (IMPQ) project enabled the PRDH to continue efforts to reconstitute the French Catholic population. The PRDH and BALSAC collaborated to combine Quebec parish register data across four centuries in the new infrastructure, the IMPQ database. This collaboration drew upon the 17th-, 18th- and 19th-century RPQA longitudinal data, the 19th- and 20th-century BALSAC marriage data, and Catholic births and deaths provided by IGD. Subsequently the PRDH and IGD supplemented the IMPQ data with non-Catholic marriage acts.

This article describes the content and design of the RPQA database, detailing the structure of this relational database and the breadth of variables available for data management and analysis. It then describes features of the IMPQ infrastructure, focusing on the extension of family reconstitution from 1800 to 1849, describing observational protocols necessary to use these data and investigating data completeness after 1800. This article should be read in conjunction with our 2018 *History of the Family* publication which provides further information on the origins of the PRDH, issues posed by migration, missing data, record linkage procedures and research possibilities (Dillon et al., 2018). Finally, we address a fundamental question, "What is my population?" as part of a broader reflection upon the target population encompassed by these data. Pushing family reconstitution forward to the mid-19th century and working with complete-count census data has forced the PRDH team to reckon with an increasingly mixed and geographically mobile Quebec population which included both Catholics and non-Catholics. Fortunately, collaboration with genealogical partners have enhanced our purchase on the diverse ethnic groups and enabled us to broaden our mandate beyond the settled French-Catholic population.

2 BACKGROUND AND SCOPE OF THE RPQA DATABASE

The PRDH was launched by Hubert Charbonneau and Jacques Légaré in 1966. At the time it was one of a handful of university-based programmes to computerize historical population data (Charbonneau et al., 1967). The PRDH family reconstitution efforts were facilitated by the exceptional circumstances of historical parish register preservation in Quebec: Catholic parish registers were maintained from the beginning of the colony and annual copies of each parish register were duplicated for civil authorities, ensuring that a comprehensive set of parish records were preserved for posterity (Bouchard & LaRose, 1976; Desjardins, 1998). The PRDH created its own microfilm images of Quebec's parish registers prior to 1700, and then turned to microfilmed images created by the Genealogical Society of Utah (Desjardins, 1998, p. 216; LaRose, 2015, p. 172) as well as a different set of microfilm images created by IGD (LaRose, 2015, p. 171). The PRDH conducted all transcription of the 17th- and 18th-century parish acts, as well as death acts after 1800 of persons born before 1750 (for an example of these sources, see Figure 1).

An article published at the very beginning of the project conveys an ambitious and comprehensive objective, stating that

[...] we attempt to establish the most complete file possible of all individuals who migrated to or were born in the specified territory (Quebec, where one finds a relatively complete series of parish registers) in the course of a particular period. The selection will extend from 1608 to 1850 [...] (Charbonneau, Légaré, Durocher, Paquet, & Wallot, 1967, pp. 216–217, translation).¹

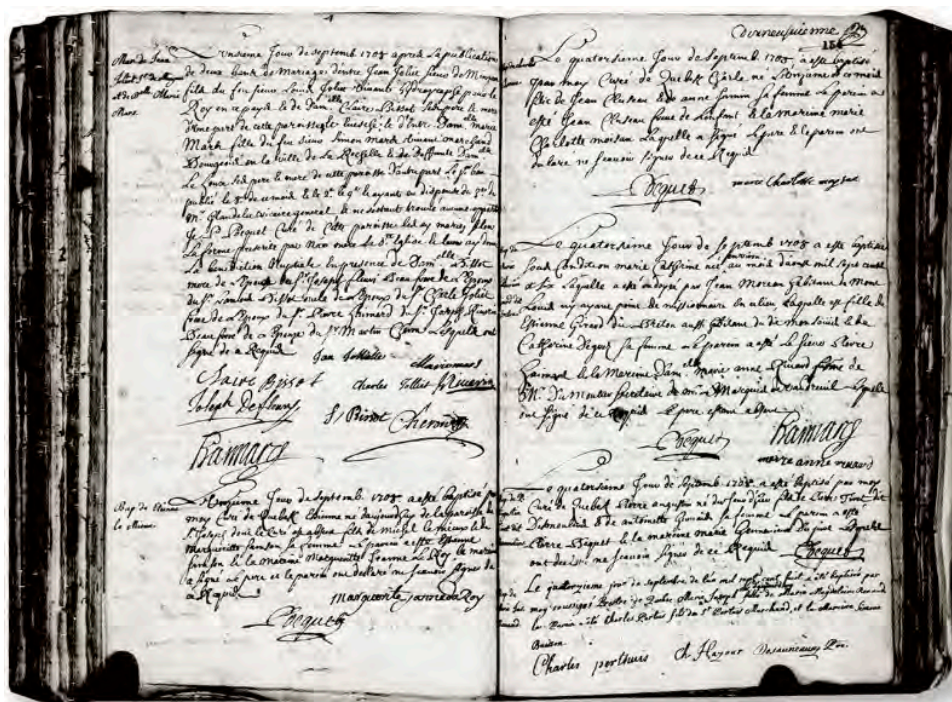
1 «Pour ce faire, nous tenterons d'établir un fichier aussi complet que possible de tous les individus venus ou nés dans un territoire défini (le Québec, où se trouve une série relativement complète des registres paroissiaux) au cours d'une certaine période. La cueillette s'étendra de 1608 à 1850 [...].»

For the past 50 years, the PRDH maintained a 100% record linkage philosophy which aimed to reconstruct, as exhaustively as possible, all demographic events for a complete population. As explained at length in our 2018 article:

When an act is missing, we not only glean information from other acts in order to infer events and dates, but, more importantly, we construct identities for persons who would otherwise remain ghosts in the documentary corpus [...] The pursuit of fragmentary lives and the completion of individual and family biographies is an important part of producing as complete a file as possible. (Dillon et al., 2018, p. 12).

Accordingly, the project incorporated into the database all of the Catholic parishes which lined both sides of the St. Lawrence River and which comprise the Quebec colony then known as "Le Canada". The PRDH devoted extra attention to missing events and dates in the 17th and 18th-century, filling in gaps by consulting complementary sources such as censuses and notarial acts (Desjardins, 1993). It exerted efforts to identify the founders, or heads of ascending genealogical lines, who first immigrated into Quebec, as well as, in some cases, the French parents of these founders who never set foot in Quebec (persons identified as "Hors Population"). It also traced persons who left the St. Lawrence Valley to conduct mission work or pursue the fur trade in places such as Détroit and Michilimackinac, insofar as Catholic parish registers are available for those locations. The PRDH drew the line, however, at persons who did not appear in the Catholic parish register collection, in other words, who did not integrate into the French Catholic population, notably non-Catholic indigenous persons and Protestants. The Quebec colony is often characterized as a "semi-closed population", with limited in- and out-migration after the initial period of settlement up to 1700, although some outmigration from the colony did occur at certain times (see Dillon et al. (2018) for an extended discussion of migration issues concerning the RPQA). Over time, the PRDH integrated into the database references from the *Mémoires de la Société généalogique canadienne-française*, the publication *l'Ancêtre*, notarial acts, the *Dictionary of Canadian Biography* and religious community archives in order to identify 1,074 persons who left the colony and approximate their date of departure. Emigrants are identified in the database by the Emigrant variable, while particular records pertaining to outmigration are identified by the Type_Acte variable. While a third of these departures occurred before 1700, about 10% occurred during the 1750s and a quarter during the 1760s. Researchers using the RPQA benefit from these efforts by the PRDH to identify outmigrants combined with consistent observation of inter-parish migrants, and are thus better equipped to control for migration selection bias.

Figure 1 *Registre de Notre-Dame-de-Quebec, including one marriage act and four baptismal acts, September 3 and 4, 1708*



Source: Image provided with permission by ©2016 Institut généalogique Drouin.

3 THE RPQA DATABASE

3.1 STRUCTURE AND VARIABLES AVAILABLE FOR ANALYSIS

The RPQA is stored in a 24-table Microsoft SQL relational database held on a dedicated server hosted by the Université de Montréal. The principal variables in this database are described in Appendix 1. The transcription of each individual act is recorded in the ACTE and MENTION tables. Transcribed acts are mainly baptismal, marriage and burial acts, but other kinds of acts are also present in the RPQA. Over 60,000 complementary historical sources have also been integrated into the RPQA database, namely the 1666, 1667 and 1681 censuses of the colony, the 1716 and 1744 censuses of Quebec City, marriage contracts, hospital sick lists, and lists of migrants. Other types of acts which have been added to the RPQA include naturalizations, testimonies of freedom to marry, recantations, confirmations, marriage rehabilitations (post-hoc legitimizations of unions contracted by related persons or conducted by a civil authority), and marriage annulments.

The ACTE table contains one line for each act which denotes basic information pertaining to the act (see Appendix 1). The variable "idActe" is a unique number assigned to each act, and once assigned, this number is never overwritten or re-used. The variable "Type" describes 20 different act types, from baptisms, marriages and burials to censuses, fur trade contracts, researcher-contributed information, marriage contracts and several other types. The date of event variable ("DateEvenement") describes the date of the demographic event (such as a birth) while the date of registration variable ("Date") describes the date of registration of the demographic event (such as a baptism). A date of registration is always available but for 20% of acts between 1621 and 1849 the date of the demographic event is missing; in these cases, the researcher must rely upon the date of registration of the event. When the date of the demographic event is known, over 90% of event registrations occurred within three days of the event itself, with a quarter occurring on the same day. The gap between dates of birth and dates of baptism fluctuated on the basis of the dispersion of the population, the number of priests, the relative influence of priests within their community, the climate, the day of birth, and urban-rural status, with greater delays apparent among rural and winter births (Amorevieta-Gentil, 2010, p. 110). On the other hand, an unknown event date may signal a sizable time delay between the event and the registration of that event.² A code for the parish of registration is also recorded in the ACTE table. For parish acts dated from 1621 to 1849 and currently included in the RPQA database, the parish code specifies 331 different parishes, most of them located within the St. Lawrence Valley.³ Some parish codes actually denote hospitals (e.g. Hôpital general de Montréal) or early missions.

About 9,634 acts registered in France from 1621 to 1849 as well as 458 acts registered in other parts of Europe (notably Ireland, Germany and England) or "At sea" have been added to this corpus, mainly to identify the origins of immigrants or parents of subjects residing in Quebec. Another 59,731 acts are registered in 77 different parishes, states or localities located outside the St. Lawrence valley but within the Americas, notably from Catholic parishes in present-day Ontario and Acadie, and ranging from Port-Royal to New Orleans. Finally, as of 2022, 45,100 parish acts from 181 Anglican, Congregational, Presbyterian and Methodist churches have been added to the RPQA. These acts are mainly Protestant marriage acts contributed by IGD to help identify mixed Catholic-Protestant couples who subsequently baptized or buried their child in the Catholic church; we also observe 80 Jewish marriage acts from Montréal's Shearith Israel Congregation. In just 1% of cases the parish is unspecified. A "Provenance" variable is included in table ACTE to describe the source of the act, while a "ObiitOndoiement" variable indicates if an emergency baptism has been performed or if a birth act includes a marginal notation indicating that a death occurred (such as in the case of stillbirths). A "Consanguinite" variable identifies consanguineous marriages explicitly denoted as such by the priest. However, not all consanguineous marriages are identified in this way and researchers interested in these marriages should apply programs designed to identify various degrees of consanguinity to capture all such cases.

Each individual person named in the act is considered a "mention" and their information is recorded in separate successive lines in the MENTION table (See Appendix 1). On average, there are five mentions per act; the number of mentions per act varies based on the type of act and the number of

2 For an extended discussion of missing data, events which were not recorded at all, see Dillon et al. (2018), pp. 8–10.

3 In this paragraph and the next paragraph, the counts of 1621–1849 parish acts across different regions are drawn from the most recent version of the RPQA dated December 2021.

witnesses present at the event. Since there are an average of 4–5 mentions per act, the MENTION table is quite large: the 2,235,082 acts from 1621 to 1849 include a total 8,987,409 mentions. Each named individual is identified via a unique identification number for their specific mention (idMention), as well as a unique identification number for the whole database (idIndividu). Once attributed, the idMention number stays the same and is never re-used. The idIndividu number is also never re-used. If the identities of two individuals bearing two different idIndividu numbers are subsequently merged, the idIndividu associated with the greatest number of mentions is kept and the second idIndividu number is archived, never to be re-used. For all mentions drawn from acts from 1621 to 1849, 80% are identified with a specific idIndividu, while 20% have a value of –1 in the idIndividu variable. The –1 value indicates that our record linkage program was unable to identify this person within the family reconstitution. Such individuals include persons who spent only a limited time in the colony and seemingly never joined in a family life in Quebec; it also includes witnesses whom we were unable to identify. On the other hand, over 90% of subjects, parents of subjects and children of subjects mentioned in an act have been successfully identified and connected into the family reconstitution.

Mentions are also identified in terms of their role as subject, spouse, father, mother, son or daughter ("Role") and their status as present at the event, absent, living, deceased or unknown ("Presence"). Often absent or deceased family members, notably parents and spouses from previous marriages, are mentioned in baptismal, marriage or burial acts. As a result, the "Presence" variable can optionally be used to identify that last recorded observation of an individual in the database when their death act is absent or can be used in conjunction with a preceding act to interpolate a date of last observation. The age and marital status of each mentioned individual, if given in the act, are also recorded here, along with the first and last name as recorded in the act as well as standardized versions of each derived from the name dictionaries. Finally, about 14% of mentions state the occupation of the subject (usually male) or his or her father. Most "habitants" or farmers do not have an occupation specified; on the other hand, nobles, seigneurs and other elites are usually identified via their occupation or an honorific title, allowing researchers to distinguish elite and non-elite men. All occupations have been coded with OCCHISCO, a North Atlantic Population Project code which is an adaptation of the original HISCO coding scheme (Roberts, Woollard, Dillon, Ronnander, & Thorvaldsen, 2003). The OCCHISCO coding scheme proved adaptable to the French- and English-regime parish register data, with the creation of additional codes to represent habitants (61119) and seigneurs (20500). Slaves were coded as 99150 (Worker not further specified) as their specific tasks were usually unspecified. The PRDH create two new OCSTATUS codes to represent the statuses slave (14) and bourgeois (53).

The ACTE and MENTION tables are linked to each other via the "idActe" variable. Adding the "Date" variable from the ACTE table to the MENTION table has the advantage of transforming the MENTION table into an event file. By sorting the MENTION table on idIndividu and then Date, researchers can assemble all mentions pertaining to each individual in chronological order. There are an average of 15 dated mentions per individual with a valid "idIndividu" in the MENTION table (1621–1849), with a third of individuals having 20 or more dated mentions. These mentions will usually begin with that person's first observation in a baptism and may end hastily with their death as an infant (in such cases, the birth and the death are declared in the same act; 11% of all identified persons from 1621 to 1849 have only one mention in the MENTION table). On the other hand, the series of mentions for a particular individual may continue for several decades through marriages and remarriages, the baptisms, marriages and early deaths of their children and ending (optimally) with their death. Researchers should keep in mind that the death mention for a given individual may not be their last dated observation, as parents and spouses can be mentioned in acts pertaining to their children or widowed spouses after death. Examining the MENTION file in this way will also reveal persons for whom no clear beginning or ending observation is readily apparent.

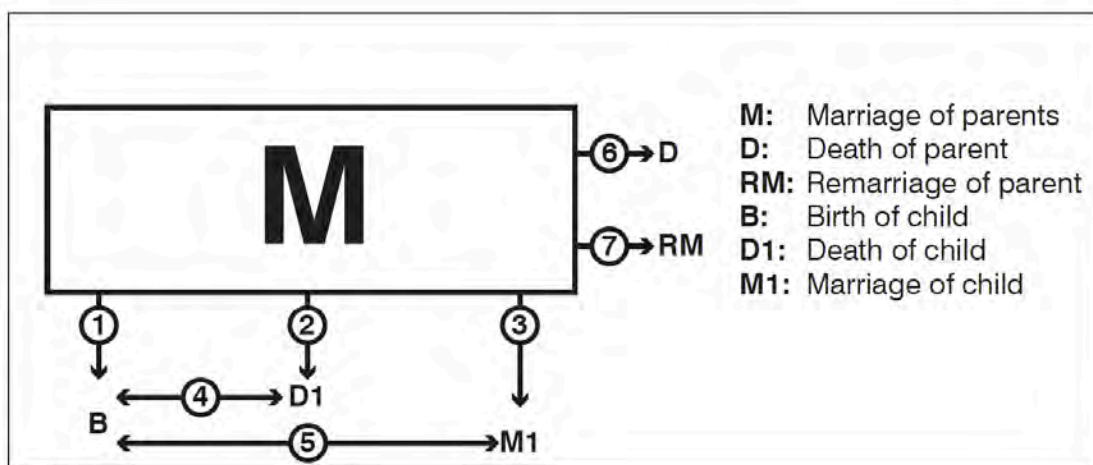
A further table named ACTE_COMMENTAIRE features a listing of over 280,000 comments either transcribed directly from the act or annotations added by a researcher or record linker. Many of these comments, notably indications that a child was baptized in emergency or that a child was born out of wedlock, or else dispensations accorded in the case of consanguineous marriages, were used to create variables in the ACTE table during manual record linkage. Other comments, read together, constitute a veritable social history of Quebec: indications of twin births, adultery ("commerce illicite"), wetnursing ("en nourrice"), abandoned or fatherless infants ("enfant anonyme"), slaves and slave owners ("panis ou panisse appartenant à ..."). The comments also feature causes of death, although only 5% of death acts cited a cause of death, usually in circumstances such as drowning, sudden death, epidemics, sickness, accidents or war.

3.2 FAMILY RECONSTITUTION

The ACTE and MENTION tables contain the core information transcribed from historic parish registers. Two other tables, INDIVIDU and UNION, are created as a result of the family reconstitution process, and are used to build our research files. Since the reproductive period for historic Quebec families is quite long, the PRDH assembles transcriptions of parish register acts spanning a period of 25 years before initiating the family reconstitution process for that block of time. Figure 2 presents the different steps taken in the PRDH Family Reconstitution Process. This process starts by opening a family file for each couple based on the marriage acts. The record linkage programme then examines systematically and in chronological order all acts which name the same couple, from the birth of their first child (item 1 in Figure 2) to the marriage or death of their last child (item 2 and 3). The quality of the Quebec Catholic parish registers, which usually included parents' names on acts pertaining to children, facilitates this family reconstitution process. The deaths referred to in item 2 concern in particular the deaths of children before marriage; in these cases, their parents are mentioned on the death act. At the same time, two further types of acts which mention parents as "subject-spouse" are also linked to the family file. These remaining acts concern the death act for each spouse (item 6, a death occurring in adulthood) and, if available, a remarriage act for the remaining spouse (item 7). "Conducting record linkage in chronological order optimizes the process: as information is chronologically cumulated, the confirmation of links becomes easier." (Dillon et al., 2018).⁴

Once all acts pertaining to children are linked to their parents' family file, a second phase of linkage creates individual biographies by linking each child's birth, marriage and death record (see record linkage items 4 and 5, Figure 2). "This second stage is aided by the fact that the pool of acts is now constrained to those previously united for the family of origin." (Dillon et al., 2018). About three-quarters of the PRDH family reconstitution is effected via automatic record linkage, using an extensive first- and last-name dictionary. The PRDH has also employed extensive manual record linkage to resolve remaining cases. Most cases linked manually are resolved easily, distinguishing between two candidates or confirming a one-to-one link by identifying an error in name or date transcription or by detecting an error in a previous linkage. The most difficult cases, such as missing marriages, Protestant marriages or marriages which are incomplete in terms of the identification of parents or previous spouses, are resolved thanks to complementary information, such as the names of witnesses or godparents or through the use of notarial acts. There are no formalized rules used by the PRDH in manual record linkage; instead, the PRDH relies on the expertise of three record linkers with extensive experience in genealogy and who have worked with parish registers for several decades.

Figure 2 PRDH Family Reconstitution Process



4 Further information on the record linkage process, including the treatment of remarriages, the automatic linkage programme, blocking, the use of a name dictionary and the creation of individual biographies via horizontal linkage, is available in Dillon et al. (2018).

Most persons listed in the baptismal, marriage and burial acts are ultimately linked into the family reconstitution. Once connected to their parents' family file, subjects are then attributed an individual identity via the assignment of an individual identification number (ID_Individu) and are listed in the INDIVIDU table. As stated earlier, mentions of individuals who are not successfully linked into the family reconstitution remain in the MENTION table with the value –1 for their ID_Individu. The INDIVIDU table features one line per identified subject and summarizes their biographical and personal information, providing the date and place of birth and death ("DateNaissance", "CodeLieuNaissance", "DateDeces", "CodeLieuDeces") and the identification numbers for each person's mother and father ("idMere" and "idPere"). The identification numbers idMere and idPere can change if a link is subsequently modified in the course of updating the longitudinal file. Data quality codes for the date of birth, "QualiteDateNaissance", and date of death, "QualiteDateDeces" indicate if the date represents the date of demographic event, the date of registration, a date from an emergency baptism, a date deriving from information from a researcher, an inferred date or missing. The INDIVIDU file also includes complementary information on immigrant and emigrant status derived from parish acts or inferred based on an assessment of the known facts about the individual ("Immigrant", "Emigrant"). The variable "Illegitime" indicates if the individual was born out of wedlock; this status is often repeated at marriage, although some illegitimate children were subsequently legitimized via the post-birth marriage of their parents. The "Amerindien" variable indicates if the individual was an indigenous person. Some persons identified in our family reconstitution never set foot in Quebec but are nevertheless mentioned as parents on marriage or burial acts. The variable "HorsPopulation" identifies these persons, enabling researchers to set them aside from analysis. The presence of the pointer variables "idMere" and "idPere" allow researchers to attach characteristics of parents to their family members, as well as establish inter- and intra-generational links.⁵

A quarter of identified men and 16% of identified women born between 1621 and 1765 contracted two or more marriages. The percentage of women and men who remarried was higher during the 17th century, particularly in the case of women on account of the unequal sex ratio and smaller number of women on the marriage market (see Charbonneau, Desjardins, Légaré, & Denis, 2000, p. 116). To accommodate multiple marriages, marriage information is stored separately in the UNION table and linked to the INDIVIDU table via the "idIndividu" variables. The UNION table contains one line per marriage act, and indicates the unique number of the union ("idUnion"), the individual identification numbers (idIndividu) of the bride and groom ("idHomme" and "idFemme"), the date and place of the marriage ("Date" and "CodeLieu"), and the data quality of the date of marriage ("QualiteDate"). The identification number idUNION is never overwritten or re-used. Since the numbers idHomme and idFemme point to the idIndividu of the spouse, they can change if a link is subsequently modified. Importantly, this file is named the UNION file because some marital unions have been inferred indirectly via information from other acts (notably children's baptisms) rather than directly via an act of marriage. This inferential process resembles our process for creating identities for persons who were never physically present in the colony but were nevertheless mentioned in acts as absent parents. As explained in our 2018 article, "The pursuit of fragmentary lives and the completion of individual and family biographies is an important part of producing as complete a file as possible" (Dillon et al., 2018, p. 12).

4 RPQA RESEARCH FILE, DISSEMINATION PROCEDURES AND NEW DEVELOPMENTS

Today, the RPQA ACTE file includes almost 640,000 original baptismal, marriage and burial acts from 1621 to 1799 (Table 1, see section 5.1). This count of baptisms, marriages and burials is somewhat less than the "true" count of births, unions and deaths which actually took place and which are recorded as events in our INDIVIDU file. By augmenting our data with complementary information such as marriage contracts, inferring missing unions from baptisms, and inferring missing deaths from remarriages, almost 750,000 births, marriages and deaths from 1621 to 1799 are available in the

⁵ The idMere and idPere variables are analogous in function to the momloc and poploc variables in the IPUMS, NAPP and PRDH historical census microdata. See IPUMS "Family Interrelationships" (<https://usa.ipums.org/usa/chapter5/chapter5.shtml>) and IPUMS-NAPP "Constructed Family Interrelationship Variables" (https://www.nappdata.org/napp/family_interrelationships.shtml), accessed January 18, 2021; Dillon, 2000.

INDIVIDU and UNION files for analysis (statistics not shown). The burial acts of 40,879 individuals born before 1750 and who died after 1800 have also been added to the RPQA, ensuring complete observation and permitting a broader demographic study of this population (Desjardins & Dillon, 2008). Contributions by researchers and genealogists have also helped the PRDH augment the family reconstitution and add variables to the database; examples include Protestant-Catholic marriages prior to 1800, listings of French soldiers from the Seven Years War 1755–1760 and African-origin and indigenous slaves. The RPQA family reconstitution features 474,000 individual biographies and 74,000 family files encompassing four or five generations and up to nine generations in certain cases. Upon request from researchers, the PRDH distributes a research file of the RPQA. The most recent version of this file is dated December 2021 and includes many corrections to pre-1800 observations made in the course of recent infrastructure projects. The research file is based on the INDIVIDU file, but also includes information from the UNION file for ease of research. Usually, the PRDH distributes the entire research file to researchers, but occasionally the PRDH prepares a particular extract of the file for those working on specific years, places or population sub-groups. For researchers who plan to conduct event-history analysis, the PRDH will also provide the MENTION and ACTE files. The PRDH has created two date variables to indicate or estimate the first and last observation of each individual. The dates of first and last observation are ideally defined via the dates of birth and death, with post-death events excluded, but for a minority will concern other events. In the case of this minority, if the last dated mention of an individual is from an act in which their presence (and vital status) is not indicated, we interpolate a date of last observation between the date of this act and a preceding act in which the individual's presence is clear.

The creation of historical census data and the integration of census and parish register data is another important part of the PRDH mandate. The PRDH created a complete-count database of the household-level 1831 Census of Quebec (78,049 household heads) and linked half of these household heads to the family reconstitution data (Cherkesly, Dillon, & Gagnon, 2019). For the same project, the PRDH transcribed over 450,000 occupations from the 1800–1824 birth and death acts as well as 46,000 occupations from the birth and death acts of the 1825–1849 parishes of Montreal, Quebec City, Trois-Rivières, Gaspésie and Saguenay/Charlevoix, ensuring researchers using Quebec family reconstitution data can integrate socio-economic status into their analyses. These occupations have been coded with OCCHISCO. Most recently, the PRDH is working with *The Canadian Peoples/Les populations canadiennes* project to code and prepare complete-count datasets of the 1851 to 1921 Canadian census microdata. The resulting infrastructure, including the PRDH's data file of the 1881 Canadian census, created in collaboration with FamilySearch, will include 40 million observations (Baskerville & Inwood, 2020, p. 598).

In December 2020, the PRDH obtained new funding for the project *Transcending borders: A historical demographic infrastructure for the study of French-Canadian families in motion*. This project will pursue Quebec family reconstitution up to 1861, incorporating Protestant acts within Quebec and Catholic acts in cross-border Ontario parishes. The Institut Généalogique Drouin has contributed to the project over 350,000 marriage acts from 1800 to 1861, drawn from its Connolly collection, as well as over 880,000 baptismal and burial acts from 1850 to 1861. Access to the IGD Connolly marriage acts, which carry no data usage restrictions, will facilitate data management as well as data legacy protection at the PRDH. The IGD is also contributing to the project Catholic birth, marriage and burial acts from bordering communities in Ontario, namely the counties of Glengarry, Stormont, Prescott and Carleton (Ottawa). Linking Ontario Catholic records into the database will enable researchers to explore the lives of francophones who moved back and forth across the Quebec-Ontario border. A renewed collaboration with FamilySearch as well as with Ancestry will enable the PRDH to integrate complete-count census observations into the longitudinal data, beginning with the 1831 and 1852 censuses and eventually incorporating the 1825, 1844 and 1861 censuses. The PRDH will then close the family reconstitution file with linkage to 19th-century nominal Canadian censuses. Integrating census observations into longitudinal data offers an unbiased way to establish which individuals were still present in the population as of the date of the enumeration and thereby maximize usage of the data (Gutmann & Alter, 1993). The PRDH offers historical data files and access to the RPQA online repertoire free of charge to researchers and students; since 2015, the PRDH has responded to 70 requests from graduate students and Canadian and international researchers interested to use the PRDH data.

5 PUSHING FAMILY RECONSTITUTION FORWARD: ANALYSIS OF THE IMPQ FAMILY RECONSTITUTION, 1800–1849

5.1 THE IMPQ DATABASE

The recent extension of Quebec family reconstitution forward to 1849 was accomplished via an inter-university and inter-sectoral collaboration. In 2006, the PRDH first entered talks with IGD to plan the transcription of baptismal and burial acts from 1800 to 1849; the IGD conducted this work from 2009 to 2011, and provided copies of these acts to the PRDH in exchange for links between the acts. The contribution of these transcribed acts from the IGD became part of the inter-institutional and inter-sectoral collaboration *Infrastructure intégrée des microdonnées historiques de la population du Québec (1621–1965)* (IMPQ) project (2013–2017), funded by the Canadian Foundation for Innovation (CFI) (Vézina & Bournival, 2020; Vézina, St.-Hilaire, Bournival, & Bellavance, 2018). One goal of this project was to extend family reconstitution to 1849, using 167,868 BALSAC 1800–1849 marriage acts (linked intergenerationally) and 1,357,899 IGD 1800–1849 baptismal and burial acts. The IMPQ project in turn drew upon an earlier PRDH-BALSAC collaboration in which *Registre de la Population du Québec Ancien (RPQA)* marriages from 1621 to 1799 were linked to the BALSAC 19th- and 20th-century marriage database to create a research file of intergenerationally-linked marriages from 1621 to 1965. Family reconstitution from 1800 to 1849 for the IMPQ project was conducted by the PRDH at the Université de Montréal, using the previously-linked BALSAC marriage records as a core set of family files into which the PRDH integrated birth and death records, adjusting marriage links and adding new unions as required. Fortunately, the PRDH was able to start the project with several tools already in place: a 24-table relational database structure, extensive first- and last-name dictionaries to facilitate name comparisons, family reconstitution protocols, an automatic record linkage program developed and refined over the course of several decades, links to over 40,000 deaths after 1800 of persons born prior to 1750 and access to Quebec Protestant marriage records from IGD.

The 1800–1849 family reconstitution conducted as part of the IMPQ project was closed in November 2018, when a copy of the INDIVIDU, UNION, ACTE and MENTION tables containing longitudinal data from 1621 to 1849 were delivered to the BALSAC office at the Université du Québec à Chicoutimi, whose server is used to house the IMPQ files. All 17th- and 18th-century data included in the IMPQ file was derived from the RPQA. The 1621–1799 family reconstitution data as well as the 1800–1849 extension were contributed with the agreement that the inter-institutional title *IMPQ* would be used rather than a single-institution brand. As a result, the title and citation to be used by researchers is: "*Integrated Infrastructure of the Quebec Population Historical Microdata (1621–1965)* (IMPQ) [Database]. Université du Québec à Chicoutimi/Université du Québec à Trois-Rivières/Université de Montréal/Université Laval." The tables were transferred with PRDH identification numbers (idIndividu, idUnion and idActe) permitting researchers to establish a correspondence between the research file disseminated by the IMPQ project and the RPQA infrastructure resident in Montreal. Researchers can request access to the IMPQ database via the IMPQ website, <https://impq.cieq.ca>, or by sending a message directly to impq@uqtr.ca.

The new family reconstitution data from 1800 to 1849 includes 1,596,526 original baptismal, marriage and burial acts (Table 1). The PRDH linked these data to preceding RPQA acts, and the resulting 1621 to 1849 file includes 2,235,082 acts.

Table 1 *Distribution of baptism, marriage and burial acts, by period, Quebec 1621–1849*

Period	Baptisms	Marriages	Burials	Total
1621–1699	21,307	3,857	6,166	31,330
1700–1759	116,125	20,402	65,274	201,801
1760–1799	239,821	39,377	126,227	405,425
1800–1824	316,087	104,943	158,291	579,321
1825–1849	616,128	103,614	297,463	1,017,205
1621–1849	1,309,468	272,193	653,421	2,235,082

Source: RPQA & IMPQ. File data_ACTE.2019-01-09.sav.

Explanation: No selections applied.

5.2 ASSESSMENT OF THE 1800–1849 IMPQ DATA COMPLETENESS

To facilitate research using the 1800–1849 family reconstitutions in the IMPQ database, we present a brief assessment of data completeness.⁶ Since family reconstitution is a long-term process, requiring a period of at least 75 years to complete the observation of most persons, and since our family reconstitution is not yet closed with a nominal census, it is important to understand the extent of data completeness in the last decades of the study period. Tables 2 and 3 show the completeness of this family reconstitution, indicating the percent of individuals with linked births, unions and deaths. Table 2 focuses on the period 1621 to 1824, selecting persons who were born or married or died before 1825. By selecting persons with a birth and/or a marriage and/or a death before 1825, Table 2 identifies persons who were included in our family reconstitution well in advance of the end of our study period in 1849. In Table 2, the number of individual biographies which include a birth linked forward to one or more marriages and then to a death (the optimal scenario for the study of fertility) is 157,307, or 19% of the file from 1621 to 1824. Another 274,246 individuals have a known date of birth linked to a date of death but no known marriage; such persons are 34% of the whole file, many of whom would be children who died in infancy. Finally, 24% of individual biographies feature a date of birth linked to one or more marriages, but not to a corresponding date of death; these would be individuals born in the early 19th century who died after the end of our family reconstitution (197,600 persons). Another 16% of individuals are in a similar situation, born during the 19th century but have not yet either married or died before the end of the observation period (130,898 individuals). From these case counts, we see that there are 629,153 individuals with a beginning observation (birth date between 1621 and 1824) and with either a marriage and/or a death to close observation. These cases represent 77% of all biographies.

The rules of historical demography discourage the use of marriages to close observation, but since the Quebec population exhibited high marriage intensity, marriages can optionally close observation for certain research topics such as infant mortality, especially if the observation period is delineated in such a way that record linkage coverage reaches high levels. Researchers desiring to study a population bounded by clear birth and death dates will have access to 431,553 cases from 1621 to 1824, 53% of the biographies (see percentages Table 2A). Researchers who wish complete birth-to-death biographies which represent the majority of the population can opt for a shorter observation period based on individuals born from 1621 to 1760. In this case, the researcher will have access to 139,116 cases with birth and death dates out of 164,696 biographies, or 84 % (see Table 3).

Table 2 *Completeness of individual biographies, RPQA & IMPQ microdata, 1621–1824*

Frequency Distribution					
	No Date of Union		Date of Union		TOTAL
	No date of death	Date of death	No date of death	Date of death	
No date of birth	11	10,871	46,300	951	58,133
Date of Birth	130,898	274,246	197,600	157,307	760,051
TOTAL	130,909	285,117	243,900	158,258	818,184
Percentage Distribution					
	No Date of Union		Date of Union		TOTAL
	No date of death	Date of death	No date of death	Date of death	
No date of birth	0.00	1.33	5.66	0.12	58,133
Date of birth	16.00	33.52	24.15	19.23	760,051
TOTAL	130,909	285,117	243,900	158,258	818,184

Source: RPQA & IMPQ. File data_INDIVIDU.2019-01-09.sav.

Explanation: The table includes all persons born in or who immigrated into Quebec including persons who emigrated from Québec, and excludes persons who never resided in Quebec ("Hors Population" status).

6 This assessment addresses the family reconstitution based on births, marriages and deaths from 1800 to 1849. The full IMPQ database also contains linked BALSAC marriages up to the 1920s (since one year of marriages is added each year, the end date of available marriages = current year - 100).

Table 3 *Completeness of individual biographies, RPQA & IMPQ microdata, 1621–1760*

Frequency Distribution					
	No Date of Union		Date of Union		TOTAL
	No date of death	Date of death	No date of death	Date of death	
No date of birth	132	668	19	169	988
Date of Birth	14,896	60,372	9,696	78,744	163,708
TOTAL	15,028	61,040	9,715	78,913	164,696
Percentage Distribution					
	No Date of Union		Date of Union		TOTAL
	No date of death	Date of death	No date of death	Date of death	
No date of birth	0.08	0.41	0.01	0.10	988
Date of Birth	9.04	36.66	5.89	47.81	163,708
TOTAL	15,028	61,040	9,715	78,913	164,696

Source: RPQA & IMPQ. File data_INDIVIDU.2019-01-09.sav.

Explanation: The table includes all persons born in or who immigrated into Quebec including persons who emigrated from Québec, and excludes persons who never resided in Quebec ("Hors Population" status).

Table 4 encompasses the period 1825 to 1849 and includes persons who were born, married or died from 1825 to 1849. This selection allows us to focus on the completeness of biographies toward the end of our study period. In this case, only 7% of biographies feature a birth linked to both a marriage and a death (70,869 cases), and 20% of biographies represent births linked directly to a death without a marriage (187,858 cases). Another 15% of biographies represent births linked to a marriage before 1850 but not yet linked to a death (144,558 cases). Less than half of the biographies toward the end of our study period include a birth linked to either a marriage or death (and just over a quarter are bounded with certainty by a birth and death).

Table 4 *Completeness of individual biographies, RPQA & IMPQ microdata, 1825–1849*

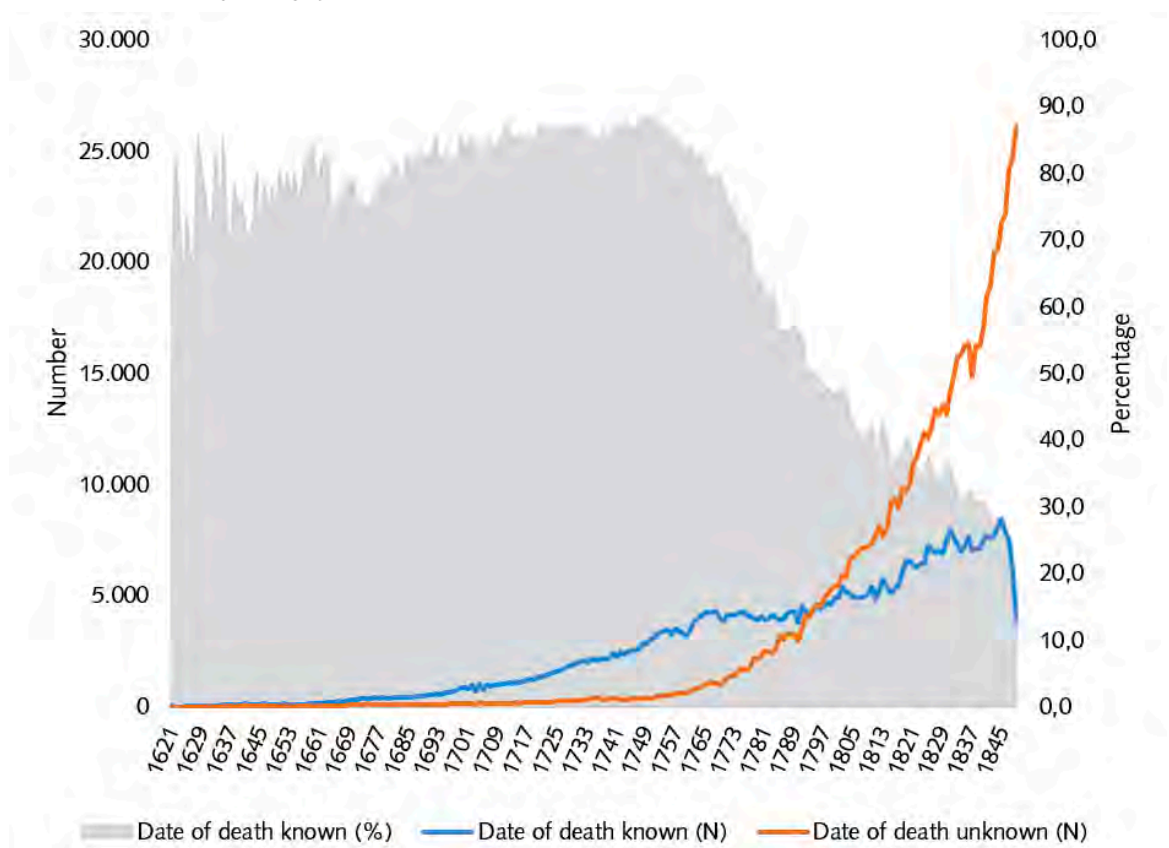
Frequency Distribution					
	No Date of Union		Date of Union		TOTAL
	No date of death	Date of death	No date of death	Date of death	
No date of birth	1	20,495	119,248	397	140,141
Date of Birth	411,151	187,858	144,558	70,869	814,436
TOTAL	411,152	208,353	263,806	71,266	954,577
Percentage Distribution					
	No Date of Union		Date of Union		TOTAL
	No date of death	Date of death	No date of death	Date of death	
No date of birth	0.00	2.15	12.49	0.04	140,141
Date of birth	43.07	19.68	15.14	7.42	814,436
TOTAL	411,152	208,353	263,806	71,266	954,577

Source: RPQA & IMPQ. File data_INDIVIDU.2019-01-09.sav.

Explanation: The table includes all persons born in or who immigrated into Quebec including persons who emigrated from Québec, and excludes persons who never resided in Quebec ("Hors Population" status).

Figure 3 addresses the completeness of individual biographies in the Quebec family reconstitution over time, showing the proportion of births linked to a known death, by year of birth, from 1621 to 1849. Obviously, the proportion of births not yet linked to a known death rises sharply toward the end of our study period as the closing of the biographies is constrained to younger and younger ages at death; 50% or more of persons born after 1794 are not linked to a known date of death (see the grey area in the graph).

Figure 3 Number and percentage of births with known date of death by year of birth, Quebec 1621–1849



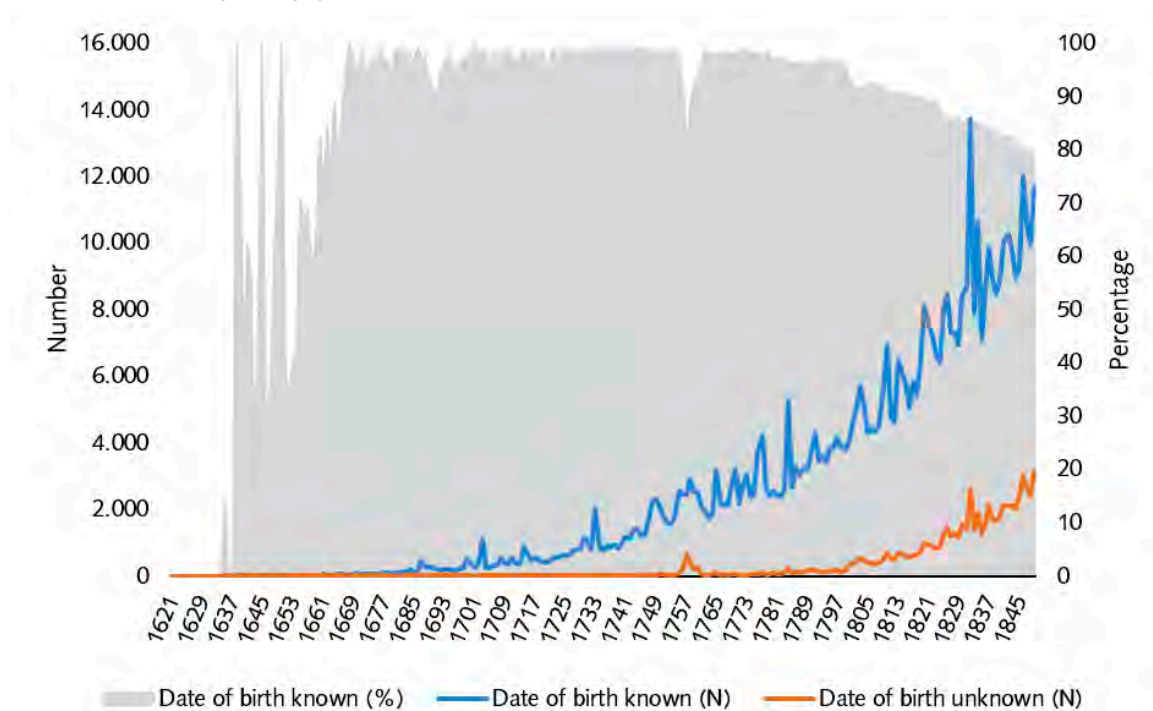
Source: File data_INDIVIDU_2019.01.09.sav.

Figure 4 presents the opposite view, the proportion of deaths in the database linked backward to a known date of birth from 1621 to 1849. This graph shows that more than 80% of individuals with a known death before 1849 and more than 90% of individuals with a known death prior to 1820 are also associated with a known birth. The proportion of deaths between 1820 and 1849 which are not yet linked to a known birth is in part due to incomplete horizontal or biographical linkage of acts from 1825–1849. Horizontal linkage of individual biographies from birth to marriage to death is always a final step after successive stages of vertical linkage (linking children's acts to the parental union), to avoid linking the wrong child baptism to the wrong child death or child marriage within sets of siblings. This hazard exists because the stock of first names among French Catholics in Quebec is relatively homogenous, and parents often re-used the same first name on a succeeding child if a prior child died in infancy. By the completion of the IMPQ project in November 2018, death and birth acts of children from 1825–1849 had been vertically linked to the parental union but many had not yet been "fused" or horizontally linked to form an individual biography.⁷ The process of vertically linking child deaths to the parental record automatically generates an inferred birth year act, and the horizontal linkage phase reconciles (or "fuses") that inferred birth year act with a genuine birth act. As a result, the RPQA/IMPQ family reconstitution file includes, in some cases, a surplus of births associated with each parental file for the 1830s and 1840s. Researchers are therefore cautioned to limit for the moment fertility and mortality analyses to births and deaths occurring before 1825.

Record linkage from 1825–1849 in the 2018 Quebec family reconstitution is also incomplete because the PRDH encountered greater linkage difficulties after 1824. These difficulties required an increasing proportion of costly manual interventions to confirm birth and death links. The PRDH also needed to consult Protestant records to identify parents and determine a marriage place and date. We present here an investigation into the proportion and type of cases which required a manual record linkage intervention, to understand the scale of this dilemma and potential bias presented by using the IMPQ family reconstitution data after 1825.

7 For further detail on the PRDH record linkage process, including vertical and horizontal linkage phases, see Dillon et al. (2018, pp. 10–12).

Figure 4 *Number and percentage of deaths with known date of birth by year of death, Quebec 1621–1849*



Source: File data_INDIVIDU_2019.01.09.sav.

5.3 MANUAL RECORD LINKAGE OF BIRTHS AND DEATHS

The PRDH approach to family reconstitution has always been based on a combination of manual and automatic record linkage. For the 1800–1849 record linkage initiative, the PRDH deployed manual record linkage tasks to employees and volunteers using a series of Excel spreadsheets, most of which have been preserved. As a result, the authors have been able to create a variable indicating if the birth or death of an individual was linked vertically to the parental union via automatic or manual record linkage. In the context of the French Catholic population of Quebec the majority of births and deaths can be linked vertically to the parental file via automatic record linkage. Table 5 shows that overall, only 13% of births and 17% of deaths required individual attention to complete the vertical link. Over time, the proportion of births linked manually increased slightly, to reach 15% for the decade 1840–1849. The proportion of deaths which required a manual intervention to complete the vertical linkage increased notably, from 14% during the first two decades of the 19th century to 19–20% by the 1830s and 1840s.

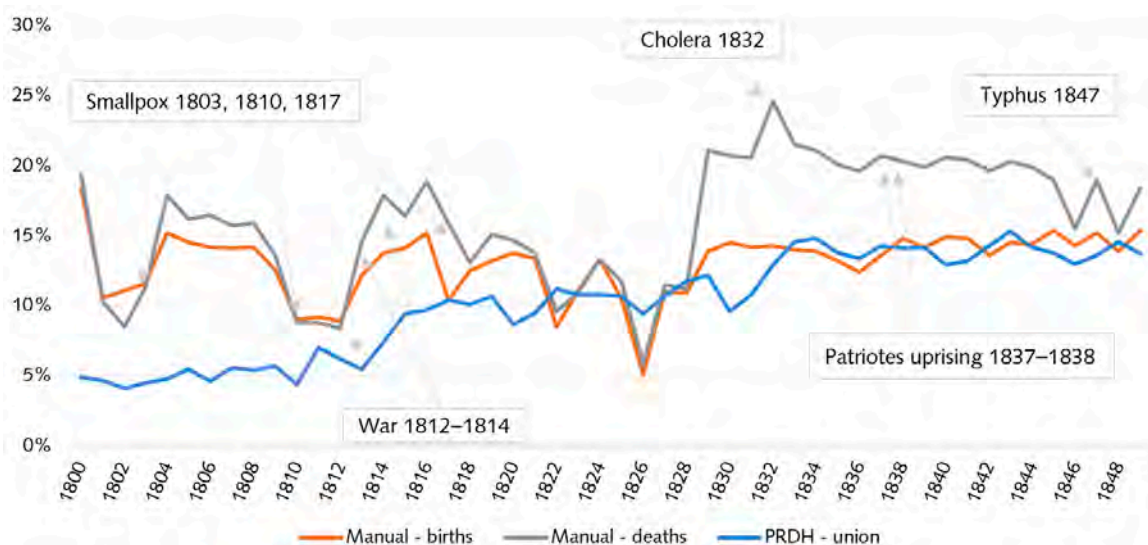
Table 5 *Percentage of births and deaths linked manually or inferred, and percentage of unions which are PRDH-inferred or Protestant unions, IMPQ microdata, 1800–1849*

	Births*	Deaths*	Unions**
%	13.3	16.8	11.4
% by decade			
1800–1809	13.7	14.5	5
1810–1819	11.9	13.8	8.1
1820–1829	11.2	12.4	10.6
1830–1839	13.9	20.9	13.3
1840–1849	14.6	18.8	13.9
N	914,428	434,371	224,297

Source: *File data_INDIVIDU_2019.01.09.sav and **UNION_2018-11-02.

Explanation: The table includes all persons born in or who immigrated into Quebec including persons who emigrated from Québec, and excludes persons who never resided in Quebec ("Hors Population" status).

Figure 5 Percent distribution of births & deaths* linked manually or inferred, and percent of unions** which are PRDH-inferred or Protestant, IMPQ microdata 1800–1849



Source: *From File data_INDIVIDU_2019.01.09.sav; **From UNION_2018-11-02 (PRDH-inferred & Protestant unions).

Explanation: The table includes all persons born in or who immigrated into Quebec including persons who emigrated from Québec, and excludes persons who never resided in Quebec ("Hors Population" status).

Figure 5 shows annual proportions of manually-linked deaths and births from 1800 to 1849, which can be divided into two periods. Before 1827, there are marked fluctuations in the proportions of manual links for both deaths and births, which rise and fall together. The proportions of deaths and births manually linked fell to their lowest levels, about 5%, in 1826. This pattern may indicate that some of the Excel work files used for manual linkage have not been preserved. Proportions of manual links for both deaths and births level off after 1827 with consistently higher interventions to link deaths than births. In this period, the proportion of deaths manually linked responds distinctly to epidemics of cholera in 1832 and typhus in 1847, but these events are not reflected in the series for births. In contrast, smallpox epidemics and war between 1800 and 1825 do not produce different responses in manual linkage for deaths and births.

Although fluctuations in manual linkages before 1827 may be due to missing work sheets, it is likely that automated linkage of deaths was less effective after 1827. Proportions of manually linked births after 1827 are roughly the same as peak years before 1827, but manually linked deaths are much more common after 1827. Our record linkers report that the quality and coherence of first name registration degraded significantly over the course of this period; Catholic priests would baptize a child with one first name and use a completely different first name on a burial act just days, weeks or months later. Any degradation in first name quality impedes automatic record linkage and increases the amount of time needed for manual record linkage. Poorer name quality should have a greater impact on linking parents' unions to child deaths than to child births, because parents' names routinely appeared in baptism records but did not appear consistently in burial records of married adults.

The diminished quality of Catholic birth and death registration after 1830 noticed by the manual linkers may be related to the increased number of parishioners per priest during this period. As the French-Canadian Catholic population kept growing rapidly in the course of the 19th century, the supply of priests probably did not at first keep pace. From 1810 to 1830, the average number of parishioners per priest rose from 1,375 to a peak of 1,834 (Hamelin, 1961, as quoted in Roy, 2001, p. 42), diminishing thereafter as the Catholic church began to open new parishes and recruit new priests.

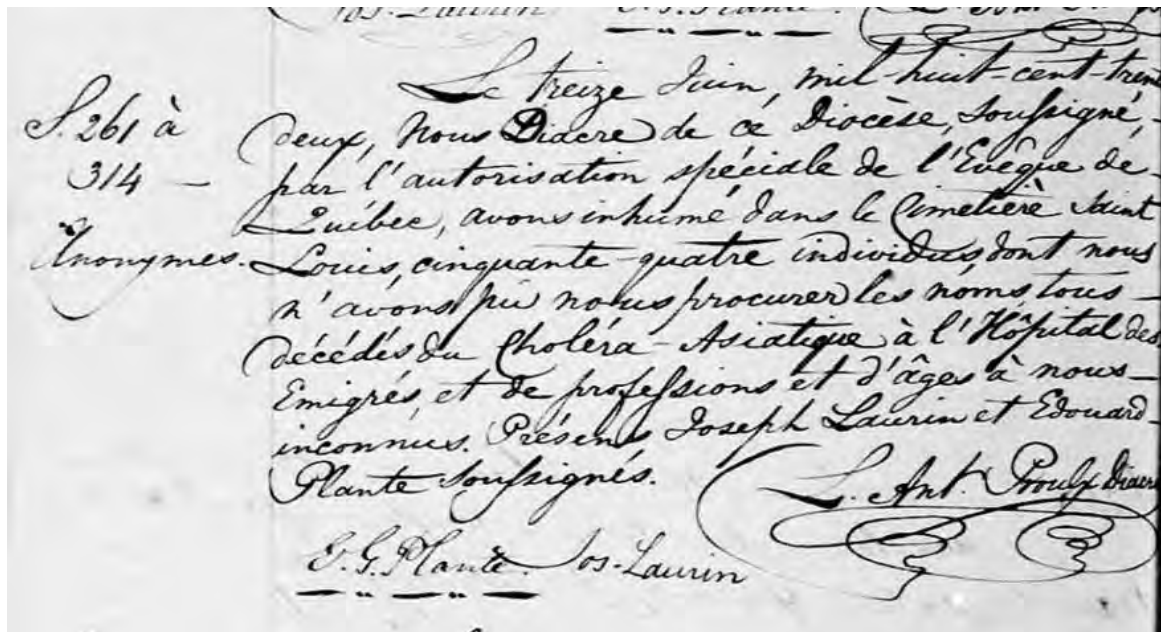
Greater record linkage difficulties might also have been occasioned by rapid social change and times of crisis. An earlier investigation of infant and juvenile mortality registration during the 17th and 18th century by Gagnon and Mazan concluded that epidemic outbreaks of diseases such as smallpox did not lead to under-registration of deaths (Gagnon & Mazan, 2009, p. 1611). However, by the 19th century, priests were handling an increasing number of new immigrants as well as rural-to-urban

migrants; the problem was not so much one of under-registration as quality of recorded information. For example, manual linkage interventions became increasingly necessary to link the births and deaths of children with 1 or 2 Protestant parents, although such cases represented only 2% of all manually-linked births and all deaths from 1800 to 1849.

Challenges encountered when manually linking records between 1800 and 1849 also suggest that when epidemics or wars occurred, the sudden rise in deaths was accompanied by a deterioration in the quality of record-keeping, including attentiveness to record both father and mother's first and last names and place of residence or place of birth. This hypothesis warrants a more focused follow-up study, beyond the scope of this paper. However, we observe that manual linkage interventions rose in the 1–3 years following the smallpox epidemics of 1803 and 1810 (but not 1817); manual linkage interventions also rose during the years of the War of 1812–1814 (Figure 5). Very telling is the way manual record linkage of deaths peaked at 25% in 1832, the year of a cholera epidemic. PRDH record linkers encountered at least 400 death acts for mass burials in Québec City during that year.

The majority of mass burials were registered in the Notre-Dame parish of Québec City from June 13 to July 1, 1832, upon the arrival of immigrant boats. Figure 6 provides an example of one such act which recorded the deaths of 54 individuals, "for whom we were not able to procure names, all having died of Asiatic-Cholera at the Emigrant Hospital, and professions and ages unknown to us." (translation by author).⁸ The sudden increase in burials may have created an administrative burden for the church; in the act shown in Figure 6, it was not a priest who wrote the act and conducted the inhumation, but rather the deacon Louis-Antoine Proulx, acting by "special authorization". The PRDH created a death act in the ACTE file for each recorded death (see Figure 7), but the lack of name and age information prevents integration of each such death into the family reconstitution. While these acts are listed in the ACTE table, no corresponding individual identity has been assigned and the death is thus absent from the INDIVIDU table. In Figure 5, we also view an uptick in manual linkage of deaths during the year of a typhus epidemic in 1847, though no appreciable increase in manual linkage during the uprising of the Patriots in 1837.

Figure 6 Example of a mass burial, Notre-Dame Parish, Québec City, June 13, 1832



Source: Image d1p_16170767.jpg provided with permission by ©2016 Fonds Drouin. <http://www.prdh.umontreal.ca/RPQA/img/acte/4341131>. Accessed January 8, 2021.

8 « [...] cinquante-quatre individus, dont nous n'avons pu nous procurer les noms, tous décédés du Choléra-Asiatique à l'Hôpital des Émigrés, et de professions et d'âges à nous inconnus. » Signed by «Diacre de ce Diocèse» Louis-Antoine Proulx.

Figure 7 *Transcription of the first mention in the mass burial, Notre-Dame parish, Québec City, June 13, 1832*



Source: <http://www.prdh.umontreal.ca/RPQA/acte/4341100/>. Accessed January 8, 2021.

5.4 IMPACT ON POPULATION PROFILE OF MANUAL RECORD LINKAGE OF PROTESTANT MARRIAGES

For the 1800–1849 marriages received from BALSAC, we do not know which unions were linked automatically or manually. However, for the purpose of this analysis, the authors have identified over 25,000 marriages which the PRDH manually integrated into the family reconstitution in one of two ways. The first case concerns baptisms or burials of children who could not be linked to a Catholic parental marriage record, but for whom one of the parents appeared to be English, Irish, Scottish, German or of another ethnicity based on last name and/or a place of origin reference within the act. In these instances, the PRDH turned to a collection of Quebec-based Protestant marriages made available to the PRDH by the IGD. The PRDH used these Protestant records to identify a date and place of the original parental union, in order to establish the family and to better specify the mother and father. In other instances of apparent Protestant-Catholic families, the PRDH could not find a corresponding Protestant marriage. Finally, the PRDH also identified child baptisms and burials for which the parents were apparently both French Catholic, but no marriage could be found. In such cases, the French Catholic parents may have married in another province, perhaps in neighbouring Ontario or New Brunswick or in an adjacent U.S. state. For these remaining acts, the PRDH used information on the children’s baptismal and burial acts to "create" the parental union. On the basis of the date of birth of the first-known child, the PRDH inferred a year of marriage for the couple. All of this work required a manual linkage intervention.

Table 5 indicates that the proportion of all unions in the file which derive from the Protestant records or which have been inferred by the PRDH is 11%. This proportion represents a minimal estimate of manually-linked marriage acts, and is an indication of the proportion of unions which would be missing if researchers rely on French Catholic marriage acts in Quebec records alone. The proportion of all unions which the PRDH needed to research in the Protestant records or which the PRDH needed to infer on the basis of information on the children’s record increased significantly, almost tripling from 5% in 1800–1809 to 14% of all unions in the 1840–1849 decade. According to Figure 5, which depicts the proportion of unions integrated into the file by the PRDH as a result of inferences based on other acts or via links to Protestant records, such links consistently rise, tracking the integration into the Quebec population of immigrants from England, Ireland and the United States. The gradual integration of British, Irish and American Loyalist immigrants into the population via inter-marriage with native-born Catholics slowed down the automatic record linkage process, requiring, over time, more and more consultation of Protestant records to link children’s acts to unions.

Family reconstitution projects such as the PRDH combine automatic and manual record linkage to improve the representativity of the population under study by maximizing links across individual biographies and between generations. With our linkage status variable, we are able to compare the population profile of persons whose births or deaths were linked manually or automatically, as well as persons whose marriage link was inferred or drawn from Protestant records (Appendix 2). The most pertinent results concern the linkage of persons who married. Table 6 (drawn from Appendix 2) shows the percent distribution of characteristics of persons included in the Quebec family reconstitution file from 1800–1849 and who contracted a first union, controlling for the linkage status (PRDH-inferred or Protestant union, BALSAC Catholic marriage act, and all persons with a first union). It is supposed that the percentages in the final column represent as closely as possible the "true" distribution of characteristics in our targeted study population.

Table 6 *Percent distribution of characteristics by union status all persons with a first union, IMPQ microdata 1800–1849*

	PRDH-inferred or Protestant	Balsac Catholic marriage acts	All unions
% Urban at first union			
% First union in Montreal, Quebec City or Trois-Rivières	53	11	16
% Other parishes	44	73	70
% Unknown	3	16	14
% Married in Catholic or Protestant parish			
Married in Quebec, Catholic parish	0	84	74
Married in Quebec, Protestant parish	97	0	12
Unknown	3	16	14
% Ethno-Religious Status at first union			
French Catholic & other Catholic	0	84	73
English-Anglican	45	0	6
Scottish-Presbyterian	17	0	2
Other Protestant	34	0	4
Unknown	4	16	15
N	50,410	356,568	406,978

Source: File data_INDIVIDU_2019.01.09.sav

Explanation: The table includes all persons with a first marriage from 1800–1849, born in or who immigrated into Quebec including persons who emigrated from Québec, and excludes persons who never resided in Quebec ("Hors Population" status).

The integration of PRDH-inferred or Protestant cases alters the distribution of characteristics of persons with a first union. The proportion of individuals with a BALSAC-linked first marriage who married in an urban place (Montreal, Quebec City or Trois-Rivières) was 11%, whereas this proportion for all persons was higher, 16%. This increase in the percentage of individuals married in an urban place is due to the integration of PRDH-inferred and Protestant unions: 53% of these additional unions took place in Montréal, Québec City or Trois-Rivières. This same result is echoed in the distribution of marriage regions. While 84% of the BALSAC-linked first unions were contracted in a Catholic Quebec parish, this proportion falls to 74% in the total population of persons with a first union, due to the integration of PRDH-inferred and Protestant marriages. The ethno-religious status of persons declaring a first union provides additional detail: we see here that 45% of persons with a PRDH-inferred or Protestant first union were married in English or Anglican churches, 17% of these unions were contracted at Scottish or Presbyterian churches and another 34% were made in other Protestant churches (mostly Methodist, Baptist or Congregationalist).

Since the integration of mixed Protestant-Catholic marriages into the family reconstitution increased the overall percentage of Protestant unions in the study population and increased the percentage of urban first unions, it is possible that the geographic characteristics of births and deaths might have been similarly affected. However, the percentage of persons who were born or died in Montreal, Quebec City or Trois-Rivières was not increased by the identification and incorporation into the file of Protestant unions (Appendix 2). In this high-fertility society, the denominators for the "Births" and "Deaths" results are overwhelmingly dominated by children born to French Catholic families. Even when we isolate births and deaths from 1840–1849, the profile of automatically-linked cases is very similar to that of the whole population (results not shown). However, the "domino" effect of incorporating Protestant unions into the database will likely be manifested to a greater extent in subsequent years, as immigration increases.

7 CONCLUSIONS AND MOVING FORWARD

Participating in the IMPQ project and pushing family reconstitution forward to the mid-19th century has motivated PRDH record-linkers to confront the increasingly mixed and geographically mobile Quebec population of the 19th century. Marital exogamy, observed in the years following the British Conquest (Angers, 2021, p. 34; Pépin, 2021, p. 35), continued during the 19th century, as a minority of French Catholic women and men partnered with Irish, Scottish and English Catholics and Protestants who migrated to Quebec. The opportunity to form mixed ethno-religious marriages varied by region across Quebec, with the greatest potential to do so in Montreal and along the borders with Ontario, New England and New Brunswick (Gauvreau & Thornton, 2015, pp. 116–117). By 1881, the proportion of mixed marriages across the province was 4.85% (Gauvreau & Thornton, 2015, p. 123); this proportion varied from 4 to 8% in Quebec City from 1852 to 1911 (Beauregard-Gosselin, 2016, p. 99; Gauvreau, Thornton, & Vézina, 2010, p. 365) and reached 13% in 1881 Gaspésie (Gauvreau, Thornton, & Vézina, 2010, p. 365). Record linkage in the Quebec context is complicated by presence of mixed Protestant-Catholic couples as well as some entirely Protestant couples who turned to Catholic priests to bury their deceased children or to baptize a new arrival. Some of these families recorded just one or two events in the Catholic registers and then disappeared; either they recorded subsequent events in Protestant registers or migrated out of Quebec. Crisis years of armed conflict and epidemics, especially in Montreal and Quebec City, coincided with upticks in the proportion of births and deaths which required manual linkage, suggesting that busy priests dealing with frequent burials took less care with names.

According to informal testimony from the PRDH record linking staff, the quality of first name registration also degraded more generally over the course of the 19th century, at least for the cases linked in a manual way, possibly in relation to the increased number of parishioners per priest and in relation to periods of war or epidemics. Identifying and incorporating this population-on-the-margins into our database required extra resources, but has resulted in a more complete file than we would have, had we relied on automatic record linkage techniques alone. These results suggest that family reconstitution of the Quebec population subsequent to 1850 will need to incorporate complete-count census records (soon to be available for the 1852 to 1921 period) into the family reconstitution process to help boost the proportion of automatic linkages. The IMPQ project has already integrated historical censuses with BALSAC civil records for the Saguenay, Gaspésie and Côte-Nord regions and for the cities of Québec and Trois-Rivières (Vézina & Bournival, 2020, p. 117; Vézina et al., 2018, p. 232 and p. 237). Notwithstanding the advantages posed by linking to censuses, the increasing outmigration of French Canadians to the central and western Canadian provinces and to the United States will pose significant obstacles for automatic record linkage. Innovative use of Canadian and U.S. complete-count census data may prove useful in that regard (see Antonie, Baskerville, Grewal, & Turcotte, 2018; Ruggles, Fitch, & Roberts, 2018, pp. 25–26). Nevertheless, family reconstitution projects will still need to set aside significant resources for manual record linkage to achieve a large percent of complete biographies and to assure representation of Quebec's increasingly multi-ethnic population. While data creators may adopt automatic linkage rules based on neutral criteria that would not apparently select the data in particular ways (for example, by avoiding the use of occupations, place of residence or secondary family members to confirm links), even the most apparently objective family reconstitution rules could lead to some form of selection if external forces such as crises, immigration and diversification of the population affect record quality and thereby linkage success rates. Continuing investment in manual record linkage as a complement to automatic record linkage, to come closer and closer to a 100% observation, would mitigate much of this potential selection bias.

To that end, the PRDH is working with genealogy partners, volunteers and graduate students to augment and complete its family reconstitution. The approach taken by the PRDH is to expand record linkage spatially (expanding beyond the Quebec border) and culturally (expanding beyond French Canadian Catholics) as well as to incorporate complementary information in the first instance, before advancing forward in time by multiple decades. This approach is preferable in order to complete and control as many individual and family biographies as possible for the purposes of historical demographic research. As a result, the PRDH continues to focus on the period prior to 1881, bringing together work on the early censuses of Quebec (1825, 1831, 1844), augmenting IGD parish records with occupation transcriptions and working on the TCP project to prepare complete-count datasets of the 1851 to 1921 Canadian censuses. The new CFI project, *Transcending Borders*, will allow the PRDH to push Quebec family reconstitution up to 1861, adding Protestant acts from within Quebec and Catholic acts from Ontario parishes across the border, and closing the reconstitution by linking to mid-

19th century censuses. This new project is thus unique for its devotion to 100% family reconstitution, focus on population closure in the period before Canadian Confederation (before 1867), pursuit of mixed Catholic-Protestant marriages and linkage to early franco-Ontarian communities. This initiative is distinct from yet complementary to the concurrent *i-BALSAC* Quebec-based historical data infrastructure project, which focuses on "a joint genealogical, genomic and geographic approach." (Vézina & Bournival, 2020, p. 117).

While the use of complete-count census records outside Quebec will help to bring Quebec family reconstitution forward in the 19th and 20th centuries, researchers also need to reflect on the question "What is our population?" Shall we confine our interest to the core French Catholic population of Quebec, or a population which includes those on the social and geographic margins who intermarried and otherwise moved in and out of the community? The latest PRDH infrastructure project represents a step in this direction. Keeping its eye on the pre-Confederation period (prior to 1867), the PRDH will leverage the resources of the IGD as well as FamilySearch to extend the longitudinal database in a way that incorporates observations of French Canadians "on the margins": those who, via geographic mobility, transcended boundaries to push into new regions and who, via marriage, engaged with Irish, Scottish, English or American newcomers. Can we expand our purview even further? Thus far, family reconstitution of Canada's Protestant population has not been undertaken because historic Protestant records are less well preserved as Catholic records, and because they often lack the first and last names of both mothers and fathers, information necessary to achieve high proportions of automatic record linkage. It is hoped that the extension of Quebec Catholic family reconstitution beyond 1849, the availability of complete-count census microdata, the greater implication of diverse genealogical partners in the production of images and data, and the advent of natural language processing and machine learning has created the conditions to address Protestant record linkage. This may only be feasible on a broad scale for the second half of the 19th century when the census enumeration of entire households at two or more decennial intervals may help to establish and confirm links across birth, marriage and death acts. Finally, we need to expand our notion of our population beyond the limits of colonial settlers to include First Nations persons as well as other non-white residents, namely Afro-Canadians. In some cases, this work will require better identification of persons already included in the database, and in others, the addition of complementary records. The principles of indigenous data sovereignty, or a consideration of "[...] the rights and interests of indigenous peoples relating to the collection, ownership and application of data about their people [...]" will also need to inform our approach (Kukutai & Taylor, 2016, p. 2). We have reason to hope that collaborations with genealogical partners, as well as inter-university collaborations and the prompting of our graduate students will open doors to expand our notion of our population and thereby our database developments.

REFERENCES

- Amorevieta-Gentil, M. (2010). *Les niveaux et les facteurs déterminants de la mortalité infantile en Nouvelle-France et au début du Régime Anglais (1621–1779)* [Intensity and determinants of infant mortality in New France and at the beginning of the English regime (1621–1779)] (Unpublished doctoral dissertation). Université de Montréal. Retrieved from <http://hdl.handle.net/1866/3944>
- Angers, S. (2021). *Marrying a redcoat: Women's experiences of marriage in the British garrison of Quebec City, 1763–1820* (Unpublished master's thesis). Queen's University, Kingston. Available from <https://www.proquest.com/dissertations-theses/marrying-redcoat-womens-experiences-marriage/docview/2561963260/se-2?accountid=12543>
- Antonie, L., Baskerville, P., Grewal, G., & Turcotte, B. (2018). Population analysis of the settlement movement in western Canada. *International Journal of Population Data Science*, 3(4), 275. doi: [10.23889/ijpds.v3i4.866](https://doi.org/10.23889/ijpds.v3i4.866)
- Baskerville, P., & Inwood, K. (2020). The return of quantitative approaches to Canadian history. *The Canadian Historical Review*, 101(4), 585–601. doi: [10.3138/chr-2020-0022](https://doi.org/10.3138/chr-2020-0022)
- Beauregard-Gosselin, I. (2016). *Intégration d'une communauté minoritaire en période d'industrialisation: Les Irlandais catholiques de la ville de Québec, 1852–1911* [Integration of a minority community during industrialization: Irish Catholics in Quebec City, 1852–1911] (Unpublished master's thesis). Université Laval, Québec. Available from <http://quescren.concordia.ca/fr/resource/NE55SVNU>

- Bouchard, G., & LaRose, A. (1976). La réglementation du contenu des actes de baptême, mariage, sépulture, au Québec, des origines à nos jours [The regulation of the acts of baptism, marriage, burial, Quebec's origins to the present]. *Revue d'histoire de l'Amérique française*, 30(1), 67–84. doi: [10.7202/303510ar](https://doi.org/10.7202/303510ar)
- 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). Cambridge: Cambridge University Press.
- Charbonneau, H., Légaré, J., Durocher, R., Paquet, G., & Wallot, J.-P. (1967). La démographie historique au Canada [Historical demography in Canada]. *Recherches sociographiques*, 8(2), 214–217. doi: [10.7202/055356ar](https://doi.org/10.7202/055356ar)
- Cherkesly, I., Dillon, L., & Gagnon, A. (2019). Creating the 1831 Canadian Census Database. *Historical Methods*, 52(2), 110–127. doi: [10.1080/01615440.2019.1567419](https://doi.org/10.1080/01615440.2019.1567419)
- Desjardins, B. (1993). Un système d'information «made in Quebec». Le registre de la population du Québec ancien [A system of information «made in Quebec». The historic Quebec population register]. In J.-P. Bardet, F. Lebrun, & R. Le Mée (Eds.), *Mesurer et comprendre: Mélanges offerts à Jacques Dupâquier* (pp. 125–136). Paris: Presses universitaires de France.
- Desjardins, B. (1998). Le registre de la population du Québec ancien [The register of the Quebec population of the past]. *Annales de Démographie Historique*, 2, 215–226. doi: [10.3406/adh.1999.1946](https://doi.org/10.3406/adh.1999.1946)
- Desjardins, B., & Dillon, L. (2008). *Étude de la population québécoise du XIXe siècle* [Study of the 19th-century Quebec population]. Social Sciences and Humanities Research Council of Canada Standard Research Grant.
- Dillon, L. (2000). International partners, local volunteers and lots of data: The 1881 Canadian census project. *History and Computing*, 12(2), 163–176. doi: [10.3366/hac.2000.12.2.163](https://doi.org/10.3366/hac.2000.12.2.163)
- Dillon, L., Amorevieta-Gentil, M., Caron, M., Lewis, C., Guay-Giroux, A., Desjardins, B., & Gagnon, A. (2018). The "Programme de recherche en démographie historique": Past, present and future developments in family reconstitution. *History of the Family*, 23(1), 20–53. doi: [10.1080/1081602X.2016.1222501](https://doi.org/10.1080/1081602X.2016.1222501)
- Dillon, L., & FamilySearch. (2019). *Complete-count database of the 1831 census of Québec* [dataset]. Programme de recherche en démographie historique, Université de Montréal (distributor) and FamilySearch.
- Dillon, L., FamilySearch, North Atlantic Population Project, & Minnesota Population Center. (2019). *Complete-count database of the 1881 census of Canada* (version 2.0) [dataset]. Programme de recherche en démographie historique, Université de Montréal (distributor).
- Dillon, L., FamilySearch, & Population et histoire sociale de la ville de Québec (PHSVQ). (2018). *Complete-count database of the 1852 census of Canada west and Canada east* (version 2.0) [dataset]. Programme de recherche en démographie historique, Université de Montréal (distributor), FamilySearch & Université Laval.
- Gagnon, A., & Mazan, R. (2009). Does exposure to infectious diseases in infancy affect old-age mortality? Evidence from a pre-industrial population. *Social Science & Medicine*, 68(9), 1609–1616. doi: [10.1016/j.socscimed.2009.02.008](https://doi.org/10.1016/j.socscimed.2009.02.008)
- Gauvreau, D., & Thornton, P. (2015). Marrying 'the other': Trends and determinants of culturally mixed marriages in Québec, 1880–1940. *Canadian Ethnic Studies*, 47(3), 111–141. doi: [doi:10.1353/ces.2015.0024](https://doi.org/10.1353/ces.2015.0024)
- Gauvreau, D., Thornton, P., & Vézina, H. (2010). Le jumelage des recensements aux mariages du fichier BALSAC: Présentation de l'approche et étude exploratoire des enfants de couples mixtes à la fin du XIXe siècle [Record linkage of censuses to marriages in the BALSAC database: Presentation of the approach and exploratory study of children of mixed couples at the end of the 19th century]. *Cahiers québécois de démographie*, 39(2), 357–381. doi: [10.7202/1003590ar](https://doi.org/10.7202/1003590ar)
- Gutmann, M. P., & Alter, G. (1993). Family reconstitution as event history analysis. In D. S. Reher & R. S. Schofield (Eds.), *Old and new methods in historical demography* (pp. 159–177). Oxford: Clarendon Press.
- Hamelin, L.-E. (1961). Évolution numérique séculaire du clergé catholique dans le Québec [Secular numerical evolution of the Catholic clergy in Quebec]. *Recherches sociographiques*, 2(2), 238. As quoted in Roy, J. (2001). Un siècle de changement religieux. In S. Courvielle & N. Séguin (Eds.), *Atlas historique du Québec: La paroisse* (pp. 40–45). Québec: Les Presses de l'Université Laval. Retrieved from <https://collections.banq.qc.ca/ark:/52327/4069821>

- Kukutai, T., & Taylor, J. (2016). Chapter 1. Sovereignty for indigenous peoples: Current practice and future needs. In T. Kukutai and J. Taylor (Eds.), *Indigenous data sovereignty: Toward an agenda* (pp. 1–22). Australian National University Press. doi: [10.22459/CAEPR38.11.2016](https://doi.org/10.22459/CAEPR38.11.2016)
- LaRose, A. (2015). Le microfilmage et la numérisation des registres paroissiaux du Québec [The microfilming and digitization of Quebec parish registers]. *L'Ancêtre: Revue de la Société de généalogie de Québec*, 41(310), 170–173. Retrieved from https://www.sqg.qc.ca/client_file/upload/L-Ancetre/Les-premieres-annees/V41-N310.pdf
- Minnesota Population Center. (n.d.). IPUMS USA. Family Interrelationships. Retrieved January 18, 2021, from <https://usa.ipums.org/usa/chapter5/chapter5.shtml>
- Minnesota Population Center. (n.d.). IPUMS NAPP. Constructed family interrelationship variables. Retrieved January 18, 2021, from https://www.nappdata.org/napp/family_interrelationships.shtml
- Pépin, K. (2021). «Les Canadiennes se sont éprises des Anglais»? Les alliances mixtes chez la noblesse canadienne après la Conquête (1760–1800) [The French Canadians fell in love with the English? Mixed alliances of French Canadian nobility after the Conquest (1760–1800)]. *Revue d'histoire de l'Amérique française*, 74(3), 31–53. doi: [10.7202/1079245ar](https://doi.org/10.7202/1079245ar)
- Programme de recherche en démographie historique. (2019 & 2021). *Registre de la population du Québec ancien* [Register of the Quebec population of the past]. [Dataset]. Université de Montréal.
- Roberts, E., Woollard, M., Dillon, L., Ronnander, C., Dillon, L. Y. & Thorvaldsen, G. (2003). Occupational classification in the North Atlantic Population Project. *Historical Methods*, 36(2), 89–96. doi: [10.1080/01615440309601218](https://doi.org/10.1080/01615440309601218)
- Ruggles, S., Fitch, C. A., & Roberts, E. (2018). Historical census record linkage. *Annual Review of Sociology*, 44, 19–37. doi: [10.1146/annurev-soc-073117-041447](https://doi.org/10.1146/annurev-soc-073117-041447)
- Trudel, M. (2004). *Dictionnaire des esclaves et de leurs propriétaires au Canada français* [Dictionary of slaves and their landowners in French Canada]. Montréal: Hurtubise.
- Vézina, H., & Bournival, J.-S. (2020). An overview of the BALSAC population database. Past developments, current state and future prospects. *Historical Life Course Studies*, 9, 114–129. doi: [10.51964/hlcs9299](https://doi.org/10.51964/hlcs9299)
- Vézina, H., St.-Hilaire, M., Bournival, J.-S., & Bellavance, C. (2018). The linkage of microcensus data and vital records: An assessment of results on Quebec historical population data (1852–1911). *Historical Methods*, 51(4), 230–245. doi: [10.1080/01615440.2018.1507771](https://doi.org/10.1080/01615440.2018.1507771)
- Vézina, H., Jomphe, M., Lavoie, E.-M., Moreau, C., & Labuda, D. (2012). L'apport des données génétiques à la mesure généalogique des origines amérindiennes des Canadiens français [The contribution of genetic data to the genealogical measurement of the Amerindian origins of French Canadians]. *Cahiers québécois de démographie*, 41(1), 87–105. doi: [10.7202/1012981ar](https://doi.org/10.7202/1012981ar)

APPENDIX

Appendix 1 *Principal variables by table*

Variable label	Variable description
<i>Table: Acte</i>	
idActe	Unique identification Number of Act
Type	Type of Act
DateEvenement	Date of Event
Date	Date of Registration
Provenance	Provenance of Act
Consanguinite	Consanguinity declaration by priest
ObiitOndoiement	Emergency baptism or marginal notation on birth act indicating a death occurred
imageTag	IGD image tag
<i>Table: Mention</i>	
idActe	Unique identification Number of Act
idMention	Unique identification Number of Mention
Role	Role of individual in the act (subject, spouse, father, mother)
Sexe	Sex
Rang	Rank of subject in the act (sequential)
idIndividu	Unique identification number of individual
Age	Age in days, weeks, months or years, as declared in act or according to a description indicating that an infant lived a few hours or a few moments
EtatMatrimonial	Marital status as declared in act
nom	Last name as recorded in act
prenom	First name as recorded in act
nomStandard	Standardized version of last name
prenomStandard	Standardized version of first name
Presence	Indication of individual as Present at event, Absent, Living, Deceased or Status Unknown
AptitudeASigner	Whether individual signed their name, did not sign their name or unknown
Profession	Occupation and/or social status
Residence	Place of residence, as described in act
Origine	Origin, as described in act (can be an ethnicity or a place)
<i>Table: Individu</i>	
idIndividu	Unique identification number of individual
idPere	Unique identification number for father (father's idIndividu)
idMere	Unique identification number for mother (mother's idIndividu)
DateNaissance	Date of Birth
CodeLieuNaissance	Code for parish of birth
QualiteDateNaissance	Quality code for date of birth
DateDeces	Code for date of death
CodeLieuDeces	Code for parish of death
QualiteDateDeces	Quality code for date of death
Sexe	Sex
Illegitime	Illegitimate at birth (born to unmarried parents)
Immigrant	Immigrant (born outside Quebec) (0 or 1)

Emigrant	Emigrant (emigrated from Quebec before death) (0 or 1)
Amerindien	Indigenous status (0 or 1)
HorsPopulation	Outside the population (never lived in Quebec) (0 or 1)
CodeOrigineEthnique	Ethnic origin
nomStandard	Standardized version of last name
prenomStandard	Standardized version of first name

Table: Union

idUnion	Unique identification number for the union
idHomme	idIndividu of the groom
idFemme	idIndividu of the bride
Date	Date of Union
QualiteDate	Quality code for the date of union
CodeLieu	Code for parish of marriage

Appendix 2 *Percent distribution of characteristics by linkage status all persons with a birth and/or death and/or first union, IMPQ microdata 1800–1849*

	Births			Deaths			First Unions		
	M	A	All	M	A	All	PRDH-inferred/ Protestant	Balsac F-Catholic	All
Sex									
Female	49	48	48	47	48	47	50	52	51
Male	52	50	50	49	49	49	50	48	49
Decade of event									
1800–1809	12	12	12	10	12	12	4	10	10
1810–1819	13	15	15	12	15	15	10	15	14
1820–1829	17	20	20	14	21	20	18	20	20
1830–1839	25	24	24	32	24	25	29	24	25
1840–1849	33	30	30	32	28	29	38	31	32
Known Events									
Birth date known				92	94	94	3	69	61
Death date known	27	34	33				2	14	13
Marriage date known	15	17	17	26	26	26			
Urban-Rural status									
Urban parish of birth (Mtl/TR/QCity)	3	2	2	2	2	2	1	12.4	12.3
Urban parish of union 1 (Mtl/TR/QCity)	2	2	2	4	3	3	41	7	12
Urban parish of death (Mtl/TR/QCity)	7	6	6	23	15	16	5	13	12
Ethno-Religious Status (from marriage1)									
French Catholic & other Catholic							0	83	73
English-Anglican							45	0	6
Scottish-Presbyterian							17	0	2
Other Protestant							34	0	4
Unknown							4	17	15
TOTAL	127,233	832,331	959,564	73,061	361,310	434,371	50,734	360,429	411,163

Source: File data_INDIVIDU_2019.01.09.sav and UNION_2018-11-02.

Explanation: Births denominator: all persons with a birth recorded from 1800–1849; Deaths denominator: all persons with a death recorded from 1800–1849. First Unions denominator: all persons with a first marriage from 1800–1849. M = Manual or inferred linkage ; A = Automatic linkage ; All = All cases.

An Overview of the BALSAC Population Database

Past Developments, Current State and Future Prospects

Hélène Vézina

BALSAC Project, Université du Québec à Chicoutimi

Jean-Sébastien Bournival

BALSAC Project, Université du Québec à Chicoutimi

ABSTRACT

The BALSAC database, developed since 1971, contains data on the Quebec population from the beginnings of European settlement in the 17th century to the contemporary period. Today, BALSAC is a major research infrastructure used by researchers from Quebec and elsewhere, both in the social sciences and in the biomedical sciences. This paper presents the evolution and current state of the database and offers a perspective on forthcoming developments. BALSAC contains marriage certificates until 1965. Coverage is complete for Catholic records (80 to 100% of the population depending on the region and the period) and partial for the other denominations. Birth and death certificates from all Catholic parishes have been integrated for the period 1800–1849 and work in underway for 1850–1916. All the records entered in BALSAC are subject to a linkage process which, ultimately, allows the automatic reconstitution of genealogical links and family relationships. The basic principle has remained the same since the beginning, namely to match individuals based on the nominative information contained in the sources. The changes made in recent years and the resulting gains are mostly related to IT advances which now offer more flexibility and increased performance. Future perspectives rest on the diversification of the sources of population data entered or connected to the database and, as a corollary, by continuous optimization of data processing and linkage procedures. In the era of 'big data', BALSAC is gradually moving from a historical population database to a multifaceted infrastructure for interdisciplinary research on the Quebec population.

Keywords: Family reconstitution, Population database, Quebec population, Record linkage, Vital records

DOI article: <https://doi.org/10.51964/hlcs9299>

© 2020, Vézina, Bournival

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

The BALSAC database, developed since 1971 at the Université du Québec à Chicoutimi, contains data on the Quebec population from the beginnings of European settlement in the 17th century to the contemporary period. These data come from the digitization of civil records and have been linked together to reconstruct families and genealogical lines over almost 400 years. Today, BALSAC is a major research infrastructure used by researchers from Quebec and elsewhere, both in the social sciences and in the biomedical sciences.

BALSAC will be entering its fiftieth year shortly. Although the formal structure ensuring its management took various configurations and names¹, the main mandates have remained centered on the preservation and development of the database and on the promotion of its exploitation by the scientific community. Technological advances have, however, brought about considerable transformations in the operations surrounding these activities, both conceptually and technically.

This paper is an opportunity to describe the evolution and current state of BALSAC and to offer a perspective on forthcoming developments. We start with a brief overview of settlement history on the Quebec territory for a better understanding of the content of the database. Next, we trace the main steps in the construction of the database and present the work underway which will enrich it substantially. Then, we provide a detailed description of the content, structure and linkage methodology. Finally, we conclude by outlining development prospects for BALSAC.

2 A BRIEF OVERVIEW OF SETTLEMENT HISTORY

The Quebec population has characteristics conducive to the construction and exploitation of a population database focused on history, demography and genealogy. One of its main features is its recent formation following European exploration, imposing well-defined temporal and geographical limits for the reconstruction and investigation of its genealogical, family and even genetic heritage.

European settlement on the Quebec territory started with the arrival of French pioneers in the early 17th century (Charbonneau et al., 1993; Charbonneau, Desjardins, Légaré, & Denis, 2000). Approximately 10,000 immigrants settled and experienced a family life in the St. Lawrence valley during a century and a half of French rule. The only period of relatively high immigration was from 1663 to 1673, when the King of France sent some 800 'Filles du Roi' to overcome the shortage of women and to encourage soldiers from the 'Régiment de Carignan' to settle in the colony (Landry, 1992). The vast majority of immigrants came from France, while most of the others originated from countries bordering France. When New France became a British colony in 1759, the population was around 70,000. There were a few thousand Aborigines and the rest were almost all French Canadians settled in the Laurentian Valley.

Following the British takeover, the French-speaking immigration virtually stopped except for a few thousand Acadians² who took refuge in Quebec following deportation by the British authorities (Dickinson, 1994). The majority of newcomers were British immigrants or Loyalists escaping the American War of Independence. It should be noted that the French population being Catholic and the English-speaking immigrants being Protestants, mixed marriages took place but remained infrequent.

In the 19th century, the French-Canadian population progressively overflowed the Laurentian corridor leading to the opening of new regions. Most of the immigration continued to come from the British

1 The research conducted with BALSAC first took place within the Société de recherches sur les populations (SOREP) created at UQAC in 1976 before becoming an interuniversity group in 1982. In 1994, SOREP became IREP, the Interuniversity Institute for Population Research, bringing together researchers from seven Quebec universities. Since 2002, BALSAC users are no longer grouped within a specific entity and the database is under the joint responsibility of UQAC, Laval University, McGill University and the University of Montreal. The BALSAC Project at UQAC manages it.

2 The Acadians are descendants of French immigrants who settled in Eastern Canada in the 17th century. In 1755, the British authorities ordered the deportation of Acadians who were dispersed in France, England and the English colonies of America. It is estimated that between 2000 and 4000 Acadians settled in Quebec.

Isles as thousands of immigrants from England, Scotland and Ireland settled in the province, most of them in the urban areas of Montreal and Quebec City (McInnis, 2000b). By 1851, Quebec had 890,000 inhabitants three quarters of whom were French Canadians.

From the beginning of the 20th century, the origins of the immigrants diversified, with many newcomers arriving from Southern and Eastern Europe (McInnis, 2000a; Piché, 2003). More recently, immigrants from Asia, South America and the Caribbean have outnumbered those from Europe (ISQ, 2019). From these various movements of immigration and settlement, results a population of some 8.5 million inhabitants, within which we find, in addition to Aboriginal communities, a French-speaking majority, an English-speaking minority and a segment made up of recent immigrants. The last two groups are mainly concentrated in the Montreal region although present in variable proportions in all regions of the province. As we will see below, the coverage of each of these groups in BALSAC depends on the characteristics of the settlement history, but also on the quality and availability of vital statistics which vary substantially across groups.

3 THE CONSTRUCTION OF THE BALSAC DATABASE

BALSAC started in the early 1970s as the project of a historian, Gérard Bouchard, who had just completed his doctorate in France where he had used the methodology developed by Louis Henry for the reconstitution of families from parish registers (Fleury & Henry, 1956). As new professor at the Université du Québec à Chicoutimi, he initiated the BALSAC project aimed at reconstructing the Saguenay–Lac-Saint-Jean (see Figure 1) population from 1837, the start of the French-Canadian settlement in the region, until 1971 using the 660,000 birth, marriage and death certificates (or acts) recorded during this period. This work completed in 1986 was the first major achievement in the development of the database.

Figure 1 *Geographical location of the cities and regions referred to in the text*



Source: Centre interuniversitaire d'études québécoises (CIEQ), Université Laval

Subsequently, the database gradually expanded to include all regions of Quebec. The name 'BALSAC' comes from an acronym made up of the first letters of the name of eastern regions of the province which constituted the very first large-scale corpus. During this period, in addition to the work conducted in historical demography and social history, a vast research program on population genetics and hereditary diseases was set up. This led to the decision to prioritize the entry of marriage certificates because of the importance attached to the genealogical approach for the exploitation of the database in the field of human genetics. This

second phase of development, extending up to 2011, added to BALSAC more than two million marriages covering all of Quebec until 1965.

Since 2010, a new phase of development is ongoing led by Hélène Vézina who has succeeded Gérard Bouchard as director. The main objectives of this new stage are 1) to add births and deaths to marriages for complete family reconstitution; 2) to adapt the database structure and linkage procedures to facilitate the connection between civil records and other types of population data; 3) to facilitate access to the database by setting up web portals.

It is in this context that BALSAC piloted, from 2013 to 2017, the creation of the Integrated Infrastructure of Historical Microdata of the Population of Quebec (IMPQ) in partnership with the Programme de recherches en démographie historique (PRDH) and the Centre interuniversitaire d'études québécoises (CIEQ) (Vézina, St-Hilaire, Bournival, & Bellavance, 2018). In the course of this project, Quebec births and deaths for 1800–1849 (more than 1.6 million records) were added to the database and civil records from BALSAC were linked to the Canadian censuses from three regions (Saguenay, Gaspésie and Côte-Nord) and two cities (Quebec and Trois-Rivières) for the period 1851–1911 (see Figure 1). Thanks to this initiative, individual life courses in BALSAC now include appearances to censuses, which contribute to filling the sometimes very long intervals among vital events and to optimize the chances of success in the various linkage operations.

Since 2019, BALSAC has been conducting a new project aimed at the creation of i-BALSAC, an infrastructure to study the Quebec population through a joint genealogical, genomic and geographical approach. The project is set around five components: integration of demographic, genetic and geographic data, development of statistical and mapping tools in order to optimize exploitation of this data and implementation of a web portal for access ('BALSAC', 2020). Through the demographic component of the project, birth and death certificates for the entire Quebec population from 1850 to 1916 (approximately six million records) will be integrated into BALSAC relying on handwritten text recognition (HTR) technology. The goal is to complete families and pedigrees in BALSAC to get as close as possible to a full population coverage and give access to omics-oriented researchers, among others, to the genealogical and familial structure up to the first decades of the 20th century.

4 DATA COLLECTION

Since the beginning of the French settlement, Catholic priests kept registers of vital events. From 1679, they were given the mandate to keep these registers in duplicate, one under ecclesiastical jurisdiction and the other, sent to courthouses, by virtue of what was to become Quebec's civil registration system (Bouchard & LaRose, 1976). This method of registration was maintained under the English Regime and continued until 1994 with the reform of the Civil Code of Quebec. Almost all the registers have been very well-preserved enabling their microfilming and more recently their digitization (LaRose, 2015).

Although they were subjected to the same civil regulation, there is a marked difference between records coming from Catholic parishes and those recorded in non-Catholic parishes (essentially Protestant) which are less rich in content making their linkage much more difficult and often impossible. For this reason, the transcription of non-Catholic records in BALSAC has not been systematic and, to date, only a few regions and periods have been processed. As we will see later, ongoing developments could help correct this situation in the coming years.

Most of the data integrated into BALSAC comes from the civil copy of the registers kept in courthouses. In the first phase of development, research assistants had to go to the courthouses in Saguenay–Lac-St-Jean region to transcribe the records on paper files and then come back to the office to transfer the data on punch cards for computer processing. In the second phase, marriage records were first entered using a microfilmed copy of Quebec registers from the Fonds Drouin. Then, with the gradual digitization of civil records, we started working with images obtained from the Directeur de l'état civil and more recently from the Bibliothèque et Archives nationales du Québec (BANQ) through collaboration agreements. Transcribed records retain a double link with the register since for the majority of them, a link to the image is created, which facilitates consultation of the original document. It is also easy to trace the act in the register as we keep information on the location (type and number of the act in the margin). Finally, during linkage operations and genealogical reconstructions, it is sometimes necessary to consult external sources (genealogy websites, registers outside Quebec, etc.), to obtain, for example, additional information on events that have taken

place outside Quebec or on the origins of immigrants. Except for the data obtained through exchanges with partners, the data entry work has entirely been done manually. However, the use of handwritten text recognition (HTR)³ now opens the way to automatically transcribing large batches of data at a lower cost and over a shorter period of time.

Although an increasingly powerful tool, the HTR process faces several challenges (Vézina, Kermorvant, Bonhomme, & Bournival, 2019). While the algorithm will 'learn' how to interpret page structure and text, any lack of uniformity constitutes an obstacle that must be addressed in some way. In addition, the larger the dataset, the greater the variability. In the i-BALSAC project, the dataset includes two million images and more than 40,000 registers. The quality of the images varies, some of them being of lower grade. This variability is also present in the languages found in the registers (French and English), in the thousands of handwriting styles, as well as in the structure of the registers that can be observed across Catholic and non-Catholic denominations.

Since each image cannot be assessed manually, the process commands an automatic page recognition prior to text recognition. This step identifies pages containing text and therefore removes blank and cover pages. Then, the process identifies how the acts are structured on the pages before reading them. The ultimate goal is to extract specific entities such as names, dates, occupations, and places. As we have access to the complete transcriptions, other entities such as honorific titles, literacy, attendance, ethnocultural group and many others could be extracted subsequently.

Fortunately, the structure of acts varies little over time, which favors a uniform collection of the information contained in the registers. Excluding godparents and witnesses, the number of participants in each type of act remains constant. A birth certificate normally contains three persons, the subject and his or her parents. There are six persons mentioned for a marriage, the spouses and their respective parents, unless it is a remarriage in which case the ex-spouse will be mentioned in place of the parents (which does not exclude that parents can also be named). Finally, depending on whether the death certificate concerns a married individual or a single individual, the act will contain two (subject and his or her spouse) or three (subject and his or her parents) mentions. Each act also contains information specific to each person such as occupation, place of residence, age (major or minor in the case of a marriage), presence, ability to sign, honorary titles, geographic origin and ethnocultural group. The frequency of these characteristics varies according to the type of event, the period of registration and the person's role. Due to the costs associated with manual data entry, some information, mostly secondary actors like godparents or witnesses, were omitted. However, the use of automatic text recognition will allow the extraction of all relevant information in the register in a near future.

The guiding principle of transcription is that the data should reflect the source as precisely as possible, including the inaccuracies and errors it contains. Names and dates are entered as they appear, despite the fact that inconsistencies in these fields may reduce the chances of a successful match. Limits due to nominative variations are taken care of by the use of dictionaries and standardization steps (see the [Data linkage section](#) below).

Any addition of data requires one or more procedures to control the quality of the information, upstream or downstream. Whether during integration or linkage operations or while validating the integrity of the database, automated queries are used to detect potential sources of errors or internal inconsistencies (see the [Validation section](#) below).

5 CONTENT OF THE DATABASE

As shown in Table 1, the database contains marriage certificates from the Quebec registers until 1965. Coverage is considered complete for Catholic records (which represent 80 to 100% of the population depending on the region and the period), but it is only partial for the other denominations. For the latter, great variability in the format of the certificates and in the information they contain (or do not contain) complicates or even makes it impossible to attempt linkage. This is why the transcription work was not systematically performed.

³ Work performed in the i-BALSAC project is conducted in partnership with Teklia (<https://teklia.com/>) and Transkribus (<https://transkribus.eu/Transkribus/>) is used for the ground truth step.

Table 1 *Spatial and temporal coverage of the three types of events in BALSAC*

	Births		Marriages		Deaths	
	From	To	From	To	From	To
SLSJ Region*	1838	1971	1838	1971	1838	1971
Charlevoix Region	1680	1945	1686	1992	1686	1992
Rest of Quebec**	1800	1849	1621	1965	1800	1849

* SJSJ = Saguenay–Lac-Saint-Jean

** Marriage records prior to 1800 come from the PRDH. They were integrated to BALSAC through a collaboration agreement.

Note: With the ongoing work as part of the i-BALSAC project, births and deaths of the whole of Quebec for the 1850–1916 period will be integrated by the end of 2022.

Birth and death certificates from Catholic parishes for the whole province are now integrated for the period 1800–1849. Only the Saguenay–Lac-Saint-Jean and Charlevoix regions (see Figure 1) have more extensive temporal coverage. Saguenay–Lac-Saint-Jean marriages, births and deaths cover the period 1838–1971, while for Charlevoix coverage extends from the end of the 17th century until the early 1990s, with the exception of births whose transcription stops in the 1940s. As mentioned above, the reconstitution of the Saguenay–Lac-Saint-Jean population represented the first phase of construction of the database. Work on the Charlevoix region, bordering and historically very connected to the Saguenay, was performed at the very beginning of the second phase before the decision to focus on marriage records was taken. Numerous studies were conducted on these two regions both from the perspective of demographic and social history and from that of population genetics (see for instance [Bouchard, 1996](#); [Bouchard & Braekeleer, 1991](#)).

Since the population database is constructed from civil records, all unregistered events are, by definition, unknown to us. In the absence of a clear denominator, it is therefore difficult to accurately measure the completeness of BALSAC coverage. Concerning the Catholic population, the rigor observed by the priests for the keeping of registers and the strict rules of transcription during the data entry process suggest that under-registration is minimal.⁴ However, some groups are clearly less well represented. We are thinking in particular of the Aboriginal groups, which have largely escaped religious registration. Although it is possible to trace individuals in the registers through mixed marriages or declarations of ethnicity or even origin, the coverage is far from optimal.⁵

This situation is, however, bound to improve. The use of HTR for the transcription of births and deaths for the period 1850–1916 will allow us to process, in addition to records from Catholic parishes, those from Protestant, Jewish and Orthodox parishes, as well as from several Aboriginal communities. Thanks to the HTR, it will also be possible to integrate the non-Catholic records of the previous periods left aside in the previous phases of development. At the end of this integration planned for 2022, the overall coverage of BALSAC and the representation of the Quebec population will thus be significantly enhanced. More than six million documents will have been processed in three years, equivalent to twice what was compiled during the first 50 years of BALSAC's existence.

The nature of the data lends itself to three levels of observation: individuals, couples (or unions) and events. The frequency of each of these units of observation and of each type of event is presented in Table 2. The number of individuals and couples are the numbers after linkage meaning that individuals and couple are counted only once independently of the number of events where they appear. It can be noticed that the number of unions exceeds the number of marriages. This is because some couples are mentioned as parents in their children's records but we do not have their own marriage certificate. Other unions are known but do not come from formal readings of Quebec register. In the context of

4 Missing data and under-registration have been extensively investigated by the PRDH researchers for the period of the French regime ([Dillon et al., 2018](#)). During the first phase of development of the database, a study was conducted at BALSAC on the parish of Laterrière in Saguenay ([Bouchard & Bergeron, 1975](#)). We also discuss this topic in a paper currently under review ([Bournival, St-Hilaire, & Vézina, 2020](#)).

5 As part of the IMPQ construction project, the censuses from the Côte-Nord, a region that contains a large Aboriginal population, display the lowest linkage rates to BALSAC, and this is particularly pronounced for the 1851 and 1861 censuses.

research projects using genealogical reconstructions, interruptions in genealogical lines are investigated in external sources and when the union is found it is integrated in the same way as a marriage. This distance taken from the registers is beneficial in that it makes it possible to add generations to lines affected by emigration or to document the reason for the interruption of a genealogical line. Over the years, tens of thousands of marriage certificates, mainly outside Quebec, were entered in BALSAC contributing to a finer understanding of migratory movements and family history of migrants.

Table 2 *Frequency of observation units in BALSAC*

Unit of observation	Number
Individuals	6,351,130
Unions (couples)	2,660,521
Events	4,327,002
Births	1,445,224
Marriages	2,303,306
Deaths	578,472

Note: With the ongoing work as part of the i-BALSAC project, the births and deaths of Quebec for the period 1850–1916 will be integrated by the end of 2022.

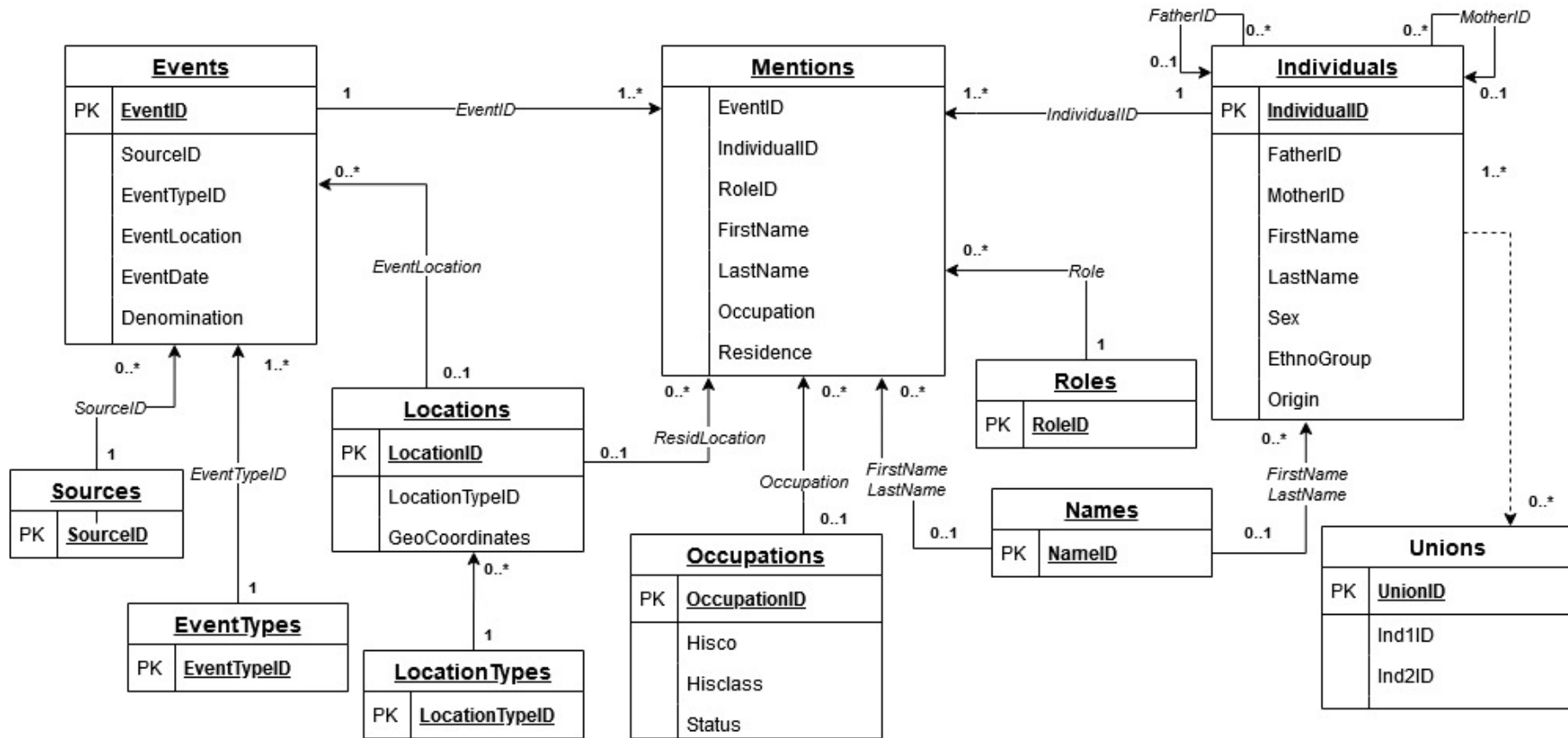
These levels or units of observation are obviously interdependent, but they can be used to answer different research questions. The observation of an individual begins with the first recorded event where he or she appears, whether as a subject or as a parent, and ends in the same way. The life cycle of a family starts when a couple gets married and ends when both members are dead or one is dead and the other remarries (unless it is lost to observation obviously).

In Quebec, data from birth, marriage and death certificates become public after 100 years. It is therefore not possible to disseminate information from vital events registered less than 100 years ago that could enable the identification of a person. The BALSAC Researchers' Support Service receives requests for access and ensures that they comply with the terms of the BALSAC Data Access Policy. It produces for researchers datasets that respect the rules of confidentiality and protection of personal information. Data recorded more than 100 years ago, which are public, are also available for consultation for research purposes on the IMPQ portal.

6 DATABASE ARCHITECTURE AND VARIABLES

BALSAC is structured as a set of relational tables where primary keys are the identifiers of events and individuals. Most of the variables come directly from the source, but some of them were created to specify or add information. For example, a date marker is used to document the accuracy of dates. In Figure 2, we focus on the content and architecture of the database according to the main source of data, namely the civil registers. But the architecture is flexible and with the integration of new data additional tables and variables can be created. The database includes three main tables *Events*, *Mentions* and *Individuals*. This is the result of recent changes made to the original structure of BALSAC in order to offer enhanced reliability as well as maximum flexibility to query the data and extract sets designed to answer the increasingly diversified needs of researchers. In fact, this modification has several advantages as it makes it possible to eliminate redundancy of information, to reduce internal consistency errors due to asynchronous information in linked tables and to develop a 'universal' linkage tool that is no longer limited to couples and can take into account other types of relationships between individuals (as described below) for the selection of candidates. The descriptive tables are devised to contain text values of entities found in each act (names, roles, occupations, locations, among others) as well as coded values that make the junction with the main tables. Another important change is that the table *Unions* no longer exists as such but rather takes the form of a dynamic view where information on couples is obtained from records in real time.

Figure 2 Architecture of the BALSAC database



Note: For the sake of concision, only the main variables of each table are presented.

The *Events* table was designed to identify and classify the different acts contained in the registers. Each event is given a unique number and characteristics such as the type of event, the place and date of registration, the date of the event as well as the source it was retrieved from are compiled. The religious denomination of the church or parish where the event was recorded is also indicated. Information on the individuals mentioned in the records are listed in two tables: *Individuals* and *Mentions*. The *Mentions* table contains all the entries likely to vary for a given person from one act to another (such as occupation, residence, role in the event, presence), while the *Individuals* table comprises the fixed characteristics such as sex, dates of birth and death, birth status (legitimate or not), geographic origin and ethnocultural group as well as the parents ID. The occurrence of these variables can vary across records, but they are always transcribed when they appear. The name of an individual can change so it is recorded in the *Mentions* table and its most frequent version appears in the *Individuals* table.

These variables are of great interest for carrying out comparative studies on various topics related to fertility, mortality and migration. It is also possible to track the social mobility of individuals and families across the occupations recorded in the events. Used in conjunction with honorary titles, occupations can serve as indicators of socioeconomic conditions. To facilitate national and international comparisons, a large part of occupations and honorary titles have been coded according to the HISCO (van Leeuwen, Maas, & Miles, 2002) and HISCLASS (van Leeuwen & Maas, 2011) classifications. Of the entire dictionary of occupations, almost 63% of the entries were coded. In terms of frequencies, the coded occupations cover 98% of all the events contained in the database. As many records contain more than one occupation, the count was done on the basis of the presence of at least one occupation coded in HISCO in the record. However, this concerns mostly men as women are clearly underrepresented in terms of the declaration of occupation (Bourque, Markowski, & Roy, 1984).

Regarding geographic variables, events generally contain several entries. First, the place of registration is always on a parish level. This is probably the most consistent information in all types of events. Then, in each type of record, certain persons (mainly subjects, parents and spouses) declare places of residence: even if the parish is often mentioned, no geographical level is prescribed so there is a great variability — from the road to the continent — which poses a problem for the standardization of the information collected. In addition to a simple ID, locations are also characterized by a specific level (*LocationTypeID*). The use of a geographic dictionary allows the different mentions to be grouped in different levels or scale units. Municipalities and parishes of Quebec are geocoded and our ongoing projects will lead to a better cartographic representation of these various levels, including census districts and sub-districts as well as some cities outside Quebec (mainly in the Canadian provinces and in the United States).

It is also possible to retrieve information on the religion, origin and ethnocultural group of individuals. The religion is not an individual characteristic per se as the information is derived from the denomination of the church where the event was recorded. As mentioned earlier, the BALSAC database covers the entire Catholic population, but only partially the non-Catholic population. The scope of this variable therefore does not depend on information contained in the registers, but rather on the source itself. Concerning origin and ethnocultural group, the information may be transcribed from the acts, but it is also searched in complementary sources when needed in research projects or genealogical work. Finally, when reconstructing genealogies, we assign a migratory status to ancestors according to whether they are native, immigrants or have never come to Quebec (most of the time the latter are parents of immigrants married in Quebec who appear in their child's marriage certificate). These variables have proven very useful in documenting the origins of the Quebec population.

7 LINKAGE METHODOLOGY

The ability to process and integrate a large number of individual data is essential when the objective is to cover an entire population. The strength of such a corpus, however, is that all of this data is linked together. All the records entered in BALSAC are subject to a linkage process which, ultimately, allows the automatic reconstitution of genealogical links and family relationships in the Quebec population.

The basis of the program has remained the same since its development in the 1970s, namely to match individuals based on the nominative information contained in the sources. The changes made since

2018 and the resulting gains are mostly related to IT advances which now offer more flexibility and increased performance. Modifications to the database architecture as well as to the name processing algorithm have also improved the efficiency of the linkage program. To put these recent developments in perspective, we first review the general principles that have helped make BALSAC what it is today.

7.1 THE GENERAL PRINCIPLES OF FAMILY RECONSTRUCTION

The family reconstitution system developed in the initial phase of the construction of the database is based on the nominative information contained in the registers and aims to group in the same file all the mentions referring to the same couple (Bouchard, Roy, & Casgrain, 1985). Thus, the basic information unit for creating the link is the 'couple mention' in a record which contains four nominative elements, namely the names and surnames of the husband and the wife as parents and as bride and groom.

The linkage program is based on two distinct and interdependent steps: the search for candidates and the linkage process itself. The mentions of candidate couples are created on the basis of at least two nominative elements common to the couple to be matched. These mentions are compared using the three modules of the FONEM phonetics program designed to detect and measure the degrees and forms of similarity between two surnames or two first names (Bouchard et al., 1985). ISG (Similarity index) calculates a score based on the degree of similarity between the nominative pairs of elements according to the position of the same letters in the names. INC (Inclusion) deals with truncated names by detecting the suffixes and prefixes in the names and deciding whether one can be treated as being included in the other. ELM (Multiple elements) deals with the situation of first and last names comprising more than one element and decides whether two entities can be treated as equivalent or not.

The linkage decision-making process is based almost exclusively on the nominative information contained in the files, but the consistency of the dates in the sequence of the family history is also taken into account (Bouchard et al., 1985). The linkage operations lead to the construction of family files which include all the events relating to a single couple (their own marriage and death, the remarriage of the surviving spouse, the births and marriages of children and the deaths of single children) as well as pedigrees with the connection of successive generations. All links created through automatic linkage and those made during the computer-assisted manual linkage stage are immediately validated by automated consistency routines (for example: acceptable interval between two events, logical sequence of events, correspondance between reported and calculated age, detection of duplicate events) (Bouchard, Casgrain, & Roy, 1981).

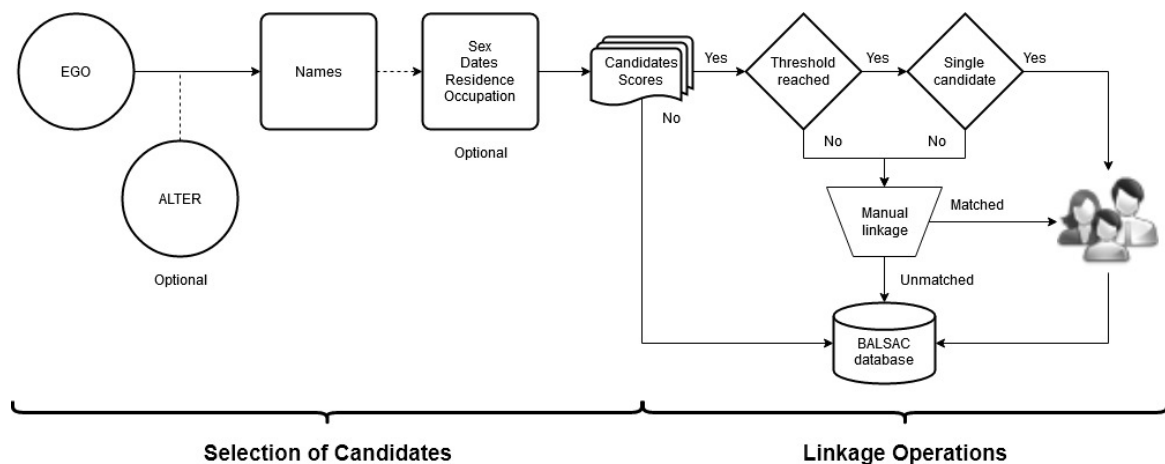
7.2 TOWARDS A 'UNIVERSAL' LINKAGE MODULE

BALSAC's current expansion goals require adapting the linkage method. As mentioned earlier, the underlying mechanics remains the same. However, we have relaxed the rules governing the linkage of civil records and introduced the flexibility necessary to open up to other types of data, while optimizing the analysis and comparison of names which are instrumental for successful linkage. The new version of the program also takes advantage of the richness of the data contained in the database such as the temporal coverage, the multiple individual occurrences, the geographic dictionary or the availability of more than one source for the same event (for instance, having both the religious and the civil copy of the register). Figure 3 presents an overview of the linkage process with its two main steps: 1) the selection of candidates which includes the calculation of a score for each proposed candidate and 2) the linkage operations where decisions are made based on a chosen threshold.

7.2.1 SELECTION OF CANDIDATES

The first adaptation aimed at improving the selection of candidates. The old version worked exclusively on the basis of couples (as parents or as bride and groom) so the selection was limited to the conjugal structure. The objective was to create a universal linkage module making it possible to search for candidates relying on various family structures by exploiting the wealth of genealogical lines. The new version of the linkage program offers the possibility of including other individuals (ALTER in Figure 3) who have a relationship with the subject (EGO in Figure 3). In the case of civil records, this mainly concerns parents, but it applies to any type of family relationship, as well as godparents and witnesses. For example, it is possible to search for a candidate named X whose father is named Y and an uncle is named Z.

Figure 3 Overview of the linkage process



Note: EGO is the subject to be linked. ALTER represents individuals related to EGO who can be used to improve candidate selection.

As they are the main keys to candidate search and matching, names remain the most important element in all operations. Previously the selection of candidates and score calculation relied exclusively on the nominative information found in marriage records. Now the program considers all occurrences or mentions of an individual to calculate a score (that is all his or her participations as subject, parent or godparent in vital events or as an enumerated individual in a census or listed in some other source). It is very common to find orthographic variations or even complete nominative variations (different names) in the list of mentions to which an individual is tied. While minor spelling variations are generally not a barrier to the selection of candidates, excessively large variations can lead to the creation of false positives or the omission of potential candidates. The program now takes into account all these variations, hence the interest in including various sources which enrich the individual data and maximize the chances of success in linkage operations.

Moreover, as the database already contains a large number of names and since all the mentions of an individual are now considered, it has become easier to link using names as they appear in registers (or any other source) without having to systematically use a standardized form like it used to be, at least for the first pass. Candidates' selection proceeds by iterations using tools called in reinforcement for the management of more complex cases when the names do not generate candidates, or not the right ones. The performance of these tools has been enhanced by the use of more sophisticated dictionaries containing orthographic variations of the same names, standardized names, patronymic equivalences, or even linguistic equivalents (names that have been translated or modified in the context of emigration, such as 'Boisvert' becoming 'Greenwood' in American censuses).

The program also offers the possibility of searching candidates according to additional variables such as dates, places or occupations. The use of these criteria is optional as they can be selected and weighted as needed. Time-changing variables are more likely to create biases in candidate selection. For residences, the calculation of a score based on the similarity of residence over the life course certainly favors sedentary individuals (same parish, same city, or even same region). This is where weights come in, making it possible to assign different weights to optional variables. For example, it is possible to decrease the weight of the declared residence in order to extend the selection and perhaps bring in candidates who would otherwise be rejected. The professions have a more limited usefulness since a large proportion of individuals will be farmers at one time or another in their life (at least up to recent times). However, a more ad hoc use targeted at professions whose occurrence is more moderate, or even low, can be an asset for the selection of candidates while risking creating a bias in favor of individuals who have a more stable professional career. It is important to consider the biases that might be introduced while using these criteria however in some contexts it might be useful and relevant to take advantage of all the information available on individuals.

All the chosen variables for candidate selection are processed simultaneously and act as blockers to discard the candidates who do not meet the search criteria. Sex is, of course, critical in this type of research since it excludes almost half of the database from the start. If a date or an interval is

mentioned, the program only retains events that occurred during this period. A slightly different treatment is assigned to the variables likely to change across records, namely names, residences and occupations. The score of these variables is an average value calculated over all the events in an individual's biography. If we consider for instance surnames, in the case of significant variations, the score will be lower since the candidate does not perfectly meet the specified criteria (it is not similar to the source in all instances). Conversely, this method ensures that no candidate who has reported the searched name at least once is eliminated.

The possibility of choosing the selection criteria and adapting the weighting system makes the program very flexible and facilitates the linkage between various primary sources and vital records. Creating models or templates makes it easier to adjust specific parameters for each data source. Also, in the context of population reconstruction, all of the available information that helps to choose between competing individuals can be mobilized during the linkage process. Thus, locations, professions and even the names of children can play a role in the calculation of scores and often lead to targeting unique candidates, which increases the possibilities of automatic linkage.

In the end, it is the summation of values for each variable that generates a final score for a specific candidate and it is this score that will determine whether there is a match or not. Obviously, the more information (in the source material and in BALSAC), the more effective the selection of candidates.

7.2.2 LINKAGE OPERATIONS

With regard to linkage, the operation is split into an automatic and a computer-assisted manual component. In both cases, the selection of candidates is involved; however, in the second, the conditions for automatic linkage have not been met and the linkage must therefore be submitted to the human eye for decision-making. For the automatic component, we run a first pass using tight criteria: a subject is linked only if there is only one potential candidate for the link and if the nominative information is perfectly identical (high score). If several individuals are proposed as candidates, a minimum threshold eliminates those whose scores are too low to justify an automatic match and linkage takes place if only if an unequivocal choice can be made (only one candidate with score above the threshold). The results of automatic linkage on the civil registers have historically hovered around 80%. As part of the IMPQ project, the linkage of censuses with BALSAC has also shown interesting results. Although the chosen method at the time was based on a human decision, we estimated that the selection of candidates would have made it possible to automatically link to the right candidate in about 75% of the cases.

A new addition to the program rests on the implementation of a second pass to increase the rate of automatic linkage. When the program has not been able to make the appropriate link with the information available in the first pass, the use of the different thesauri allows a certain level of tolerance in the face of nominal variations thus extending the selection of potential candidates. The linkage process is performed as in the first pass. Multiple candidates with scores above the chosen threshold, which are explained most of the time by homonymy (notably with Josephs and Maries) or incomplete information, and other ambiguities are referred to manual processing in proportions which depend to a large extent on source-related factors. Obviously, a certain proportion of the subjects to be linked do not exist in the database and therefore no candidates are proposed. They are then simply added to the database.

Manual linkage is a less standardized operation: it can be limited to the checking (and correcting if necessary) of the transcription by returning to the source, but it can also consist of a manual search for candidates in the database or require the use of external sources such as genealogical indexes or websites to find or clarify information and support decision making. Finally, after all these steps, there are still a number of unresolved links. As with the automatic linkage stage, these 'floating' events are kept in the database and may be the subject of further investigation. Since this can be a long and costly operation, the use of this type of inquiry is generally performed in the context of specific research needs. However, a 'silent' algorithm constantly scans the database for potential matches. Any addition of data therefore enriches the corpus and makes it possible to update certain previously unsuccessful linkages.

Finally, notwithstanding their overall quality and completeness, the registers present certain variations across time and space and, consequently, the proportion of records automatically linked will not be constant. In order to control for this and after performing various tests, we have come to the conclusion that it is preferable to 'simulate' the linkage process and to analyse the predicted results before launching

the actual automatic operations and integrate the results in the database. This process validates the consistency of the results by identifying the potential biases inferred by the data. Automatic linkage rates that appear to be too high or too low provide information on the quality of the registers or on the importance of homonymy. Different strategies can then be applied such as adjusting the parameters of the program in order to restrict or extend the selection of candidates or processing the records in selected sets in order to eliminate certain sources of noise. For instance, linkage attempted on only one parish at a time and over a defined period will allow the least mobile individuals and families to be effectively matched. It can sometimes be very wise to quickly solve these most obvious cases from the start in order to optimize automatic operations and promote gradual consolidation of the database. Subsequent operations on more difficult cases will then be facilitated.

7.3 VALIDATION PROCESS

In the development of a population database, erroneous links distort genealogical lines and may bias research results. Validation therefore represents an important step since it ensures the maintenance of the integrity of the database. A three-part process was put in place to perform data verification at crucial stages.

First, the accuracy of the transcriptions from the registers is verified. In the case of a manual entry, a research assistant validates data entry from a randomly selected sample of records. In the context of the i-BALSAC project for which the reading of millions of pages of registers has been entrusted to handwriting recognition algorithms, quality metrics assess the reading of records against information already contained in BALSAC. Recurring mentions such as surnames, first names, residences and professions can thus be standardized, despite a non-optimal reading by the algorithm. For instance, using the name dictionary, it is possible to measure the difference between a name read by HTR and the 'closest' name in BALSAC. Using these metrics, we can accept, based on a certain threshold, what seems most likely.

At the linkage stage, whether manual or automatic, integration into a family file is marked out by strict rules which ensure minimal consistency. Thus, in the call for candidates, an acceptable duration between two events is considered, in particular for intergenerational differences. In general, it is at this stage that we detect false positives and in particular problems related to homonymy. These tests aim to expose contradictions or inconsistencies in family files suspected of containing mentions referring to two distinct couples.

Even if the consistency tests can uncover the vast majority of homonymous couples wrongly linked, the nature of the data makes it inevitable that a small fraction of these cases escape these controls. In addition, since all the links specific to an individual or to a family file are not necessarily integrated in the database at the same time (for births at the end of a given period it is more than likely that death is missing), validation during data entry or linkage will not detect all internal inconsistencies. For this reason, post linkage validation of the data is also performed to identify family files that contain features that could hint to incorrect links. These checks take the form of automatic requests which can identify files containing potential duplicates, an abnormally high number of records, suspicious observation periods, sequences of improbable events (for example births too close together), inconsistencies between age declarations and actual values obtained from known dates.

8 GENEALOGIES AND KINSHIP RELATIONSHIPS

The linkage process described above has shown how generations and families are reconstituted, one record at a time. Once completed, the population database enables the reconstruction of ascending genealogies of all individuals and the exact measurement of kinship relationships between these individuals. The temporal extent of BALSAC makes it possible to trace lineages over an average of 10 generations and can extend up to 18 generations in some cases. The situation is slightly different for descending genealogies. The absence of births and deaths for certain periods does not yet allow the complete reconstitution of families, but it is possible to look at the married descendants in both extant and extinct lineages.

Using the data in BALSAC, researchers have access to individual biographies and family histories. Basic queries allow the extraction of all the kinship relations of one or more individuals up to the desired generation with the possibility of adding specific criteria (for example, relatives alive or not on a given date). From intergenerational and kinship links, it is possible to carry out several analyses which are used specifically to study genealogical datasets. A large number of queries are grouped together in a single procedure which produces, for a given corpus, the ascending genealogy of each subject, a set of descriptive measurements such as generational completeness and average depth, the portrait of the paternal and maternal lines, the list of immigrant founders, and measures of the frequency and genetic contribution of ancestors according to various characteristics (origin, sex, period of arrival, etc.). It is possible to extend these analyses in the R environment with the GENLIB package (<https://cran.r-project.org/package=GENLIB>), an open access genealogical analysis module which offers a range of functions to manage, describe and compute various measures for population genetics and genetic epidemiology (Gauvin et al., 2015)

Finally, the recent addition of godparents and witnesses from birth and marriage certificates opens the way for the analysis of extra-family networks. These persons who, mostly due to costs, had never been systematically entered into BALSAC, provide highly valuable information on social and support networks as well as their dispersion in time and space.

9 FUTURE PERSPECTIVES FOR THE DEVELOPMENT OF BALSAC

The data in BALSAC concerns mostly individuals from the past but the infrastructure that hosts and maintains it must remain anchored in the present. This is the case from a technical standpoint to keep pace with IT advances, but also from a research point of view where the needs and requests of users evolve with theoretical and methodological developments in their disciplines. One common feature is that scientists work with ever larger datasets to perform more and more complex and sophisticated analyses. The nature of the data as well as the spatial and temporal coverage makes BALSAC relevant not only for social scientists, but also for geneticists and researchers in the biomedical field. To fulfill its mission and open up a maximum of opportunities, BALSAC's offer must therefore take into account disciplinary trends and specific needs in terms of data, methods and performance.

We wish to adjust to this demand, by diversifying the sources of population data entered in the database and, as a corollary, by optimizing data processing. Various orientations have already been targeted and work is underway for some of them. We are currently working on the connection of genealogical lines interrupted by emigration. It is not uncommon for lineages to be broken off by the marriage of a couple in a locality outside Quebec, especially towards the end of the 19th century when Quebec as a whole was affected by a major wave of emigration to the United States and to other Canadian provinces. In many cases, children or grandchildren of these emigrants return to marry or settle in Quebec but there remains a gap left by the marriages outside Quebec, limiting genealogical reconstruction. Throughout the years, part of these marriages were entered in the course of research projects involving genealogical reconstruction but we now intend to proceed to their systematic integration. This represents tens of thousands of marriages recorded outside Quebec mainly in the second half of the 19th century and at the beginning of the 20th century. They come for the most part from two Canadian provinces, Ontario and New Brunswick both bordering Quebec and New England in the United States. This coverage is not exhaustive but represents a good starting point for the study of migrations, especially for the period 1840–1930.

Other information found in the registers constitutes interesting subsets to be systematically integrated to enrich life courses. We are thinking in particular of marriage licenses and documented adoptions, but also other kinds of events. Lastly, in the years 1980–1990, complementary sources such as lists of students, workers or even religious were digitized and grouped under the name 'sectoral files' without being really integrated into BALSAC. In order to build a constellation of data whose center consists of a population (and its structure) over nearly 400 years, these peripheral files are now being integrated as events in individual biographies and family histories, just like the records from the registers.

By the richness and the spatiotemporal coverage of the data it contains, BALSAC now constitutes a base on which to rely to pursue and diversify its development while continuing the integration of

data from Quebec civil records in order to get ever closer to the current period and to cover the entire population (Catholics and non-Catholics). This enrichment can be done through partnerships to bring together data of various natures such as those from historical censuses in the IMPQ or those coming from genetic research projects in i-BALSAC. Other data may be directly entered into BALSAC, for example data from the registers of French-speaking parishes in other Canadian provinces or the United States.

Other datasets of various scope and size could be connected and thus be documented by life courses and family histories. Some appear more relevant than others in the medium term, notably the causes of death, but also other documents (cadasters, directories, judicial records, hospital list, among others). These additions will enrich the family and genealogical corpus with complementary information and, conversely, provide contextual depth to the peripheral data. In the era of 'big data', BALSAC is gradually moving from a historical population database to a multifaceted infrastructure for interdisciplinary research on the Quebec population.

ACKNOWLEDGEMENTS

We express our gratitude to the Canada Foundation for Innovation, the Université du Québec à Chicoutimi and its foundation, the Université de Montréal, Université Laval and McGill University for their financial support. We also thank the research assistants, clerks and technical staff who have contributed to the development of BALSAC. We are grateful to Laurent Richard from the Centre interuniversitaire d'études québécoises at Université Laval who produced the map presented in Figure 1.

REFERENCES

- BALSAC. (2020). [Web portal]. Retrieved from <http://balsac.uqac.ca/>
- Bouchard, G. (1996). *Quelques arpents d'Amérique: Population, économie, famille au Saguenay 1838–1971*. Montréal: Boréal.
- Bouchard, G., & Bergeron, M. (1975). Les registres de l'état civil de Notre-Dame de Laterrière (1855–1911). *Archives* 75.3, 3(3), 164–173.
- Bouchard, G., & de Braekeleer, M. (1991). *Histoire d'un génome: Population et génétique dans l'est du Québec*. Sillery: Presses de l'Université du Québec.
- Bouchard, G., Casgrain, B., & Roy, R. (1981). *Tests de validation des fiches de couple par ordinateur*. Document de travail BALSAC II-C-67.
- Bouchard, G., & LaRose, A. (1976). La réglementation du contenu des actes de baptême, mariage, sépulture, au Québec, des origines à nos jours. *Revue d'histoire de l'Amérique française*, 30(1), 67–84. doi: [10.7202/303510ar](https://doi.org/10.7202/303510ar)
- Bouchard, G., Roy, R., & Casgrain, B. (1985). *Reconstitution automatique des familles: Le système SOREP*. Chicoutimi: SOREP.
- Bournival, J.-S., St-Hilaire, M., & Vézina, H. (2020). A critical assessment of historical Canadian censuses and Quebec civil registers: How linked datasets can serve as a tool to compare population microdata. *Histoire Sociale/Social History*. Manuscript under review.
- Bourque, M., Markowski, F., & Roy, R. (1984). Évaluation du contenu des registres de l'état civil saguenayen, 1842–1951. *Archives*, 16(3), 16–39.
- Charbonneau, H., Desjardins, B., Guillemette, A., Landry, Y., Légaré, J., & Nault, F. (1993). *The first French Canadians: Pioneers in the St. Lawrence Valley*. Newark, London: University of Delaware Press/Associated University Presses.
- 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). Cambridge: Cambridge University Press.
- Dickinson, J. A. (1994). Les réfugiés acadiens au Québec, 1755–1775. *Études Canadiennes/Canadian Studies*, 37, 51–61.

- Dillon, L., Amorevieta-Gentil, M., Caron, M., Lewis, C., Guay-Giroux, A., Desjardins, B., & Gagnon, A. (2018). The programme de recherche en démographie historique: Past, present and future developments in family reconstitution. *History of the Family*, 23(1), 20–53. doi: [10.1080/1081602X.2016.1222501](https://doi.org/10.1080/1081602X.2016.1222501)
- Fleury, M., & Henry, L. (1956). *Des registres paroissiaux à l'histoire de la population: Manuel de dépouillement et d'exploitation de l'état civil ancien*. Paris: I.N.E.D.
- Gauvin, H., Lefebvre, J. F., Moreau, C., Lavoie, E. M., Labuda, D., Vézina, H., & Roy-Gagnon, M. H. (2015). GENLIB: An R package for the analysis of genealogical data. *BMC Bioinformatics*, 16(1), 160. doi: [10.1186/s12859-015-0581-5](https://doi.org/10.1186/s12859-015-0581-5)
- ISQ, Institut de la statistique du Québec (2019). *Le bilan démographique du Québec. Édition 2019*. Québec. Retrieved from www.stat.gouv.qc.ca/statistiques/population-demographie/bilan2019.pdf
- Landry, Y. (1992). *Les filles du roi au XVIIe siècle: Orphelines en France, pionnières au Canada; Suivi d'un répertoire biographique des filles du roi*. Montréal: Leméac.
- LaRose, A. (2015). Le microfilmage et la numérisation des registres paroissiaux du Québec. *L'Ancêtre*, 41(310), 170–173.
- McInnis, M. (2000a). Canada's population in the twentieth century. In M. R. Haines & R. H. Steckler (Eds.), *A population history of North America* (pp. 529–600). Cambridge: Cambridge University Press.
- McInnis, M. (2000b). The population of Canada in the nineteenth century. In M. R. Haines & R. H. Steckler (Eds.), *A population history of North America* (pp. 371–432). Cambridge: Cambridge University Press.
- Piché, V. (2003). Un siècle d'immigration au Québec: De la peur à l'ouverture. In C. Le Bourdais & V. Piché (Eds.), *La démographie québécoise: Enjeux du XXIe siècle* (pp. 225–263). Montréal: Presses de l'Université de Montréal.
- van Leeuwen, M. H. D., & Maas, I. (2011). *HISCLASS: A historical international social class scheme*. Leuven: Leuven University Press.
- van Leeuwen, M. H. D., Maas, I., & Miles, A. (2002). *HISCO: Historical international standard classification of occupations*. Leuven University Press.
- Vézina, H., Kermorvant, C., Bonhomme, M., & Bournival, J.-S. (2019). i-BALSAC: Completing families with the help of automatic text recognition. In paper presented at the *Social Science History Association* (pp. 1–25), Chicago, USA.
- Vézina, H., St-Hilaire, M., Bournival, J.-S., & Bellavance, C. (2018). The linkage of microcensus data and vital records: An assessment of results on Quebec historical population data (1852–1911). *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 51(4), 230–245. doi: [10.1080/01615440.2018.1507771](https://doi.org/10.1080/01615440.2018.1507771)

HISTORICAL LIFE COURSE STUDIES
VOLUME 12 (2022), published 07-07-2022

The Ural Population Project

Demography and Culture From Microdata in a European-Asian Border Region

Elena Glavatskaya

Ural Federal University

Julia Borovik

Ural Federal University

Gunnar Thorvaldsen

UiT The Arctic University of Norway

ABSTRACT

The Ural Population Project (URAPP) is built from individual level data transcriptions of 19th- to early 20th-century parish records and mid-19th-century census-like tax revisions manuscripts. This article discusses the source material, the contents, the history of creation and the strategy of the URAPP database and the outcome of the main research topics so far, including historical demography, Jewish studies, indigenous studies and studies of religious minorities in the Urals and Siberia. Our studies of the ethno-religious cultural landscape of the Urals and northwestern Siberia as well as participation in population history projects was more vital backgrounds than the traditional focus on aggregates. The over 65,000 vital events transcribed from parish records of Russian Orthodox Churches and minority religions in and around Ekaterinburg have been the basis for studies of mortality, nuptiality, religion and other characteristics. We found that the Jewish population kept their traditions and connections with relatives in the Pale of Settlement. Prisoners of WWI usually marrying within their own religious group. Infant mortality in Ekaterinburg was lower among Jews and the Catholics, minorities with higher education and western background, while the Orthodox majority exposed their newborn to extremely tough baptism. The burial records show cases of the Spanish flu in 1918–1919, but on a lower level than in the West, supporting recent theories that estimates of flu mortality may be too high. Based on the tax revisions, polygyny was officially recognized among the indigenous Siberian people. The strategy of the URAPP project has evolved from transcribing microdata about minorities towards covering the whole population.

Keywords: Russia, Urals, Ekaterinburg, Siberia, Parish records, Censuses, Tax revisions, Nuptiality, Mortality, Ethnicity, Indigenous people, Religions, Religious minorities

DOI article: <https://doi.org/10.51964/hlcs12320>

© 2022, Glavatskaya, Borovik, Thorvaldsen

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

Today, ethno-confessional relations and demographic processes are crucial for understanding and promoting the stability and development of countries and regions. For many decades, we studied the evolution of the ethno-religious cultural landscape of the Urals with qualitative methods and sources. We now increasingly focus on social processes, involving ethnic, religious and demographic relationships and other conditions. We concentrate on Perm' province in the middle Urals during the late 19th and early 20th centuries (see map in Figure 1). The period is characterized by rapid modernization which led to changes in ethnic and religious composition, as integral parts of the first demographic transition. The traditional quantitative focus on statistical aggregates in Russia enhanced our understanding of this field only marginally. In order to understand the details of these changes, it is necessary to study the ethnic and religious communities as mirrored in source material containing information about each individual. Analysis on the individual level is also necessary to avoid ecological fallacies — i.e., drawing conclusions about smaller groups based on aggregates about society at large. In addition, in contrast to the use of published aggregate data, the method of studying selected communities at the individual level makes it possible to analyze the composition of small ethno-religious groups as parts of the larger rural and urban multi-ethnic and multi-confessional community.

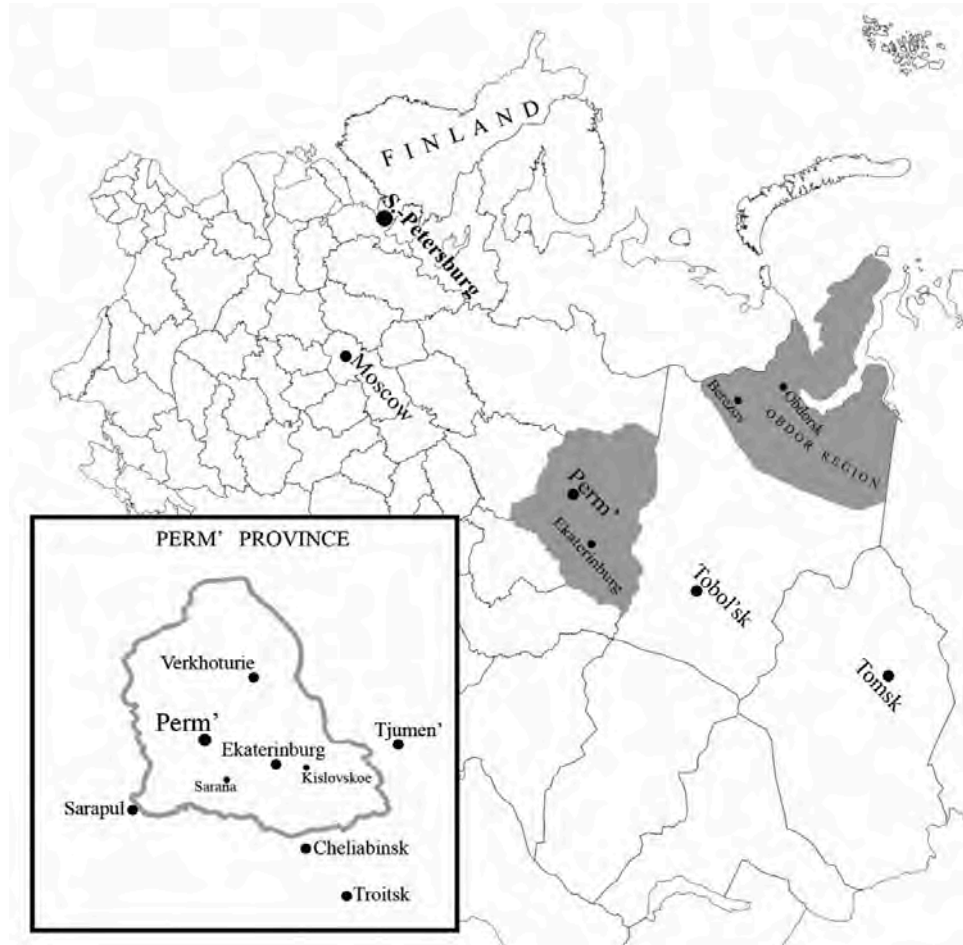
Interest in nominative microdata in Russia developed gradually during the 1980s after Heldur Palli, an Estonian historian, introduced quantitative methods in population studies in the Soviet Union (Palli, 1983). It inspired immense interest in historical demography in the 1990s due to cooperation with Western historical demographers and the introduction of computers in the Russian historians' toolkit. The Russian branch of the Association for History and Computing (AHC) founded in 1992 is still active and helpful with its professional journal *Istoricheskaia informatika* [Historical Information Science] and biannual conferences attracting both Russian scholars and researchers from the post-Soviet realm: the Republics of Belarus, Kazakhstan, Kyrgyzstan, Latvia, Ukraine and other countries (Borodkin & Vladimirov, 2017). During 30 years of new historical demography, several academic and university centers ran independent projects based on nominative data sources: in Moscow (Blum & Troitskaya, 1997; Ul'yanova & Troitskaya, 2016a, 2016b); in Sankt-Petersburg (Kashchenko & Markova, 2012; Markova, 2016); in Tambov (D'iachkov, Kanishchev, & Orlova, 2007; Strekalov & Strekalova, 2018, 2019); in Barnaul (Bryukhanova, 2019; Vladimirov & Sarafanov, 2013). Mostly, these scholars used census-like tax records and parish records.

In this contribution we describe the sources, content and strategy of the Ural Population Project dataset (URAPP). The URAPP is one of the youngest historical microdata resources, created by a group of historians at the Ural Federal University in Ekaterinburg, Russia — a city named Sverdlovsk during 1924–1991. It spans the period 1858–1959 and covers the Urals and northwestern Siberia. It consists of individual level data mainly transcribed from 18th- and 19th-century Russian census-like tax records (*revizskie skazki*) and the 19th- to early 20th-century parish records (*metricheskie knigi*) and it also includes a sample of the 1959 Soviet census. This article discusses the source material, the contents and the strategy of the URAPP database and the outcome of the main research topics so far, including historical demography, Jewish studies, indigenous studies and studies of religious minorities.

2 NOMINATIVE SOURCES IN IMPERIAL RUSSIA

While there are several types of nominative sources in Russia, we shall focus on two main ones for the following reasons: they are universal for the whole country, well preserved in Russian archives and therefore became the basis for the Ural Population Project (URAPP). They are the *revizskie skazki* (census-like tax revisions) and the *metricheskie knigi* (parish registers of vital events, hereafter called parish records), both introduced by Tsar Peter the Great (ruling from 1682 to 1725) as part of his modernization program. For an overview of other Russian and Soviet nominative sources, see (Mazur & Gorbachev, 2016).

Figure 1 Areas and localities included in URAPP database

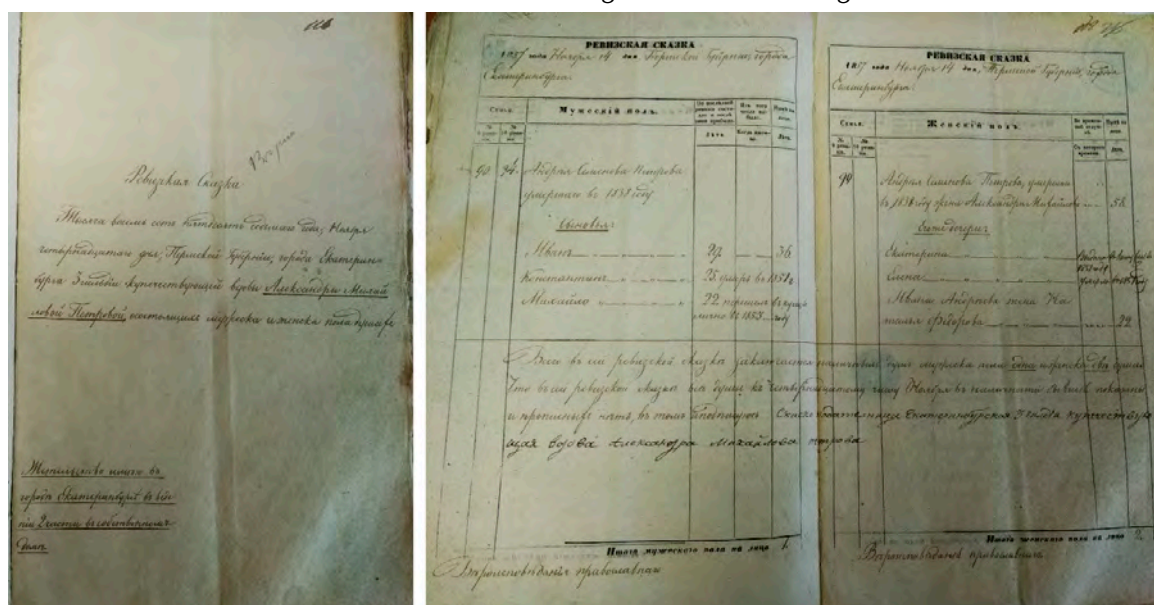


2.1 THE REVIZSKIE SKAZKI (CENSUS-LIKE TAX REVISIONS) AND CENSUSES

Even if the very concept of population registration came from the East, the Europeans modified the method and implemented modern censuses already from the 18th century (Thorvaldsen, 2018), while Russia followed only in the late 19th century (Clem, 1986). However, until the 18th century, the Russian authorities constructed lists of certain population groups on the household level, updated regularly for taxation purposes. These lists could leave out children, women, noblemen and military units, who were exempt from taxation. Table 1 provides an overview of the tax censuses that took place from 1718 to 1858. During this period 10 of them provide detailed, census-like information on individual members of the families and households (see Figure 2 for an example of these sources).

All revizskie skazki were nominative, listing the first name, patronymic and family name for the head of the household, age and social standing/status (*soslovie* [estate]) with values such as peasants, *inorodtcy* – indigenous people, *meschane* – office clerks, individual craftsmen and workers, *tcekhovye meschane* – guild craftsmen and *kuptcy* – merchants. For the rest of the family only the first name was listed, together with the relation to the head of the household and the age. New family members and data about the deceased would be registered at each revision as well as about those not present with explanation of the reason for their absence (exile, etc.). In this way the revisions combined the *de facto* and the *de jure* enumeration principles. The age information was updated according to the number of years since the previous revision, and for this reason may be inaccurate. Record linkage using records from two or more tax revisions will be facilitated by a system that numbered the households in a persistent way. The tax censuses of 1718, 1743 and 1811 excluded women, since only men were taxed. On the other hand, taxes had to be paid for the men listed until the next revision, thus also for those who had passed away in the meantime.

Figure 2 *Revizskaia skazka of the merchant's household headed by 58 years old widow Alexandra Mikhailovna Petrova living in her own house together with a son and his wife*



Source: Gosudarstvennyi Arkhiv Sverdlovskoi Oblasti [State Archive of Sverdlovskaiia oblast'], Ekaterinburg, Russia (GASO) F. 8. Op. 1. D. 1775. List 116–118.

Titles of the columns:

Left side (Males): 1. Family number according to the 9th revision, 2. Family number according to the 10th revision, 3. Name, 4. Age as registered in the previous revision, 5. Reason for absence and since what time, 6. Age at present.

Right side (Females) 1. Family number according to the 9th revision, 2. Family number according to the 10th revision, 3. Name, 4. Age as registered in the previous revision, 5. Reason for absence and since what time, 6. Age at present.

After the second revision, the authorities ordered that the revizskie skazki should be updated every 15th year. This schedule was in general followed with two exceptions: the revision of 1794 planned for 1811 was not completed due to the Napoleonic Wars and the 9th revision planned for 1850 was revamped into the 10th which started in 1856. Unlike the proper censuses taken later, it was not the aim of the revizskie skazki to mirror the population's composition on a single census day or within the timeframe of the same year for demographic purposes. We can find these lists in archives all over the previous Russian Empire. Scanned images of tax censuses for western Siberia are available at <http://archiv.72to.ru/index.php/ga-tobolsk/tobolsk-ob-material>.

Troitskaya (1995) used the revizskie skazki to construct mortality tables for the Moscow region between 1750 and 1850. In a later co-authored work (Blum & Troitskaya, 1997) the mortality estimates based on the revizskie skazki were compared with other mortality tables for the period. They also expanded the work on the Moscow region by constructing mortality tables for all of Russia during the second half of the 19th century. The authors showed that in the mid-18th century mortality in Russia was comparable to the level in France, but a century later Russian mortality remained unchanged, while the French level was significantly lower. For more information on the revizskie skazki and research made on this source, see Ul'yanova and Troitskaya (2016a, 2016b).

The first All-Russian census conducted in 1897, was carefully prepared, run and processed according to contemporary international enumeration standards. However, in accordance with what was to become Russian practice, most of the microdata were destroyed after the information was processed, aggregated and published. After this otherwise successful start of modern population census taking, there was a break due to the first Russian revolution in 1905 and yearlong insurgences making the next census impossible. The economic hardships and food crises after Russia entered World War I, required urgent information about population and supplies. Attempts to register the population in order to arrange an efficient food supply were made by Russian municipalities in 1916 and 1917. However, these and the Bolshevik attempt to take a census in 1920 partly failed, due to lack of resources during the turbulences of the foreign interventions and civil war (Thorvaldsen & Glavatskaya, 2017).

Table 1 Revizskie skazki taken in the Russian Empire 1718 to 1858

Revision number	Date of the ordinance	Period taken	Aggregate numbers
I	26.11.1718	1718–1727	15,738,000
II	16.12.1743	1743–1747	21,200,000
III	28.11.1761	1761–1767	23,200,000
IV	16.11.1781	1781–1782	28,400,000
V	23.06.1794	1794–1795	37,400,000
VI	18.11.1811	1811	41,010,400
VII	20.06.1815	1815	46,300,000
VIII	16.07.1833	1833	59,132,955
IX	01.01.1850	1850–1852	68,500,000
X	26.08.1856	1857–1858	74,556,400

Source: Andreev and Andreev (n.d.), Thorvaldsen (2018).

2.2 METRICHESKIE KNIGI (PARISH RECORDS)

From 1722 the Orthodox clergy had to perform registration of vital events in parish records. During the next two centuries, the forms evolved, and other religious denominations (Lutherans in 1764, Catholics in 1826, Muslims in 1828, Jews in 1835 and the Old Believers in 1905) were also obliged to register their vital events in the standardized state provided forms (see Figure 3 and 4).

Figure 3 Metricheskaia kniga of Ekaterinburg Synagogue on births in 1906

The image shows two pages of handwritten birth records from the Ekaterinburg Synagogue in 1906. The records are organized into columns. The first page contains entries 6, 7, and 8, while the second page contains entries 9, 10, and 11. Each entry includes the sequence number of the female (№), the sequence number of the male (№), the name of the circumciser (Имя совершавша), the year (Годъ), the day and month of birth and circumcision (Мѣсяцъ и день), the place of birth (Гдѣ родился), the social status of the father (Состояніе отца, имени отца и матери), and the name and gender of the newborn (Имя ребенка и какое ему дано или дано имя).

Source: GASO. F. 6. Op.13. D. 68. List 2–3.

Titles of the columns: 1. Sequence number of females, 2. Sequence number of males, 3. Name of the circumciser, 4. Year, 5. Day and month of birth and circumcision, 6. Place of birth, 7. Social status of the father, names of both father and mother, 8. Name and gender of the newborn.

Figure 4 *Metricheskaia kniga of Ekaterinburg Lutheran Church on deaths in 1887*

Source: GASO. F. 6. Op.13. D. 5. List 3–4.

Titles of the columns: 1. Number, 2. Date and hour of death, 3. Date and hour of funeral, 4. Christian name and family name, status, rank or occupation of the deceased; if children: First name and family name, status, rank or occupation of parents, 5. Birthplace of the deceased, 6. Age of the deceased, 7. Gender (male), 8. Gender (female), 9. Free, married, widowed or divorced, 10. Cause of death, 11. General remarks.

They all had three common parts: baptisms/births, weddings and burials/deaths. In addition, Muslim and Jewish books also had a section on divorces, which these religions allowed. This registration of vital events took place all over the Russian Empire until the October Revolution in 1917, in some places a few years longer. Russian legislation regulated the parish registers and their accuracy. The religious community board had to check and verify the books frequently and religious leaders and communities were to be fined when any disorder was found in the records (Glavatskaya & Borovik, 2019) and their quality improved significantly from the mid-19th century onwards (Mironov, 2007). After the implementation of the decree on separation of the Church from the State, a civil office established by the Bolsheviks took over the registration of vital events in each municipality, and the State did not recognize documents issued by the Church after 1917. Moreover, all the parish records from before 1917 were seized and stored in the State Archives. These collections of parish records are well preserved and together with the revizskie skazki became the source basis for the Ural Population Project (URAPP).

3 BACKGROUND OF THE URAL POPULATION PROJECT

Several independent research projects eventually brought us to the Ural Population Project (URAPP). All of them were related to ethnic or religious minorities and often dealt with Siberia. International cooperation taught us the merits of working with microdata, motivated us and prepared for launching the URAPP.

3.1 THE URAL ETHNO-RELIGIOUS STUDIES FOCUSING ON EKATERINBURG

An important step leading up to the URAPP was our multifaceted research on the evolution of the ethno-religious landscape in the Urals. A thoroughfare for migration to Siberia, the region was interesting due to its ethnic and religious diversity. It had become the meeting point for the indigenous, Muslim and Russian Orthodox traditions during the initial Russian colonization in the late 1600s, and later Catholics, Lutherans and Jews added to the religious diversity. The history of this encounter included the Christianization of the Urals and the Siberian indigenous populations (Glavatskaya, 1995, 2011a, 2011b).

Coincidentally, Swedish officers, exiled prisoners of war after the Swedish defeat at Poltava in 1709, wrote the earliest descriptions of the Siberian indigenous peoples made available in the West. We examined how the prisoners preserved their identity in exile and other aspects of their destinies by studying their diaries and other documents. It inspired our interest in religiously mixed marriages that given the shortage of Lutheran brides in Siberia, were concluded by the prisoners as an alternative marriage strategy. Many Russian widows became involved in a relationship with Swedes after the death of their husbands. However, most interfaith marriages ended when the Russian Orthodox wives were abandoned together with their children as the prisoners returned to their homeland after the 1721 peace treaty. The remaining ones, together with German contractors, made up the nucleus of the Lutheran community in the Urals with its center in Ekaterinburg (Glavatskaya & Thorvaldsen, 2015). While Lutherans were the oldest religious minority in Ekaterinburg, other denominations also added to the city's religious landscape (see Table 2).

The more than 90% Orthodox in the 1897 population census were overwhelmingly ethnic Russians, like the 4% orthodox Old Believers (Russian Orthodox who split from the mother church in the 17th century). The city's Muslim community, a religious minority of Tatars and Bashkirs, in-migrants from rural suburbs, was the third biggest. The Catholics were Polish, while the Lutherans were mainly German of origin. In addition, there were 24 Calvinists and seven Anglican Church members (likely British and Swiss), six Baptists and a Mennonite, adding to the well-established Protestant congregation. The Jews came from diverse places, mostly in western Russia. The Russian government postponed the next census planned in 1915 due to World War I, but a survey listing the names of house owners was conducted in Ekaterinburg in 1913. All the non-Orthodox denominations had expanded their share in the city, mainly due to in-migration and natural population growth; some had increased their congregation's size several times since 1897 (see Table 2).

The problem we faced when studying ethnic and religious minorities, was a lack of personal detail in the cross-sectional source material, which brought us to the parish records in the GASO archive in Ekaterinburg for the Urals region, which we shall return to in section 4 about the Ural Population Project.

Table 2 *Religious denominations in Ekaterinburg in the 1897 census and 1913 survey aggregates*

Denomination	1897		1913	
	Population	%	Population constructed	%
Orthodox	39,745	91.9	96,881	90.6
Old Believers	1,790	4.1		
Muslims	678	1.6	5,590	5.2
Protestants	343	0.8	1,245	1.2
Catholics	323	0.7	1,331	1.3
Jews	303	0.7	1,589	1.5
Other	57	0.1	251	0.2
Total	43,239	100	106,887	100

Explanation: Children under 14 were not registered in 1913. Based on the 1897 census aggregates on children, we estimated the actual population by adding 40% to each denomination in the 1913 city survey.

Source: Troinitskii (1905); GASO. F. 62. Op. 1. D. 524. List 126.

3.2 THE POLAR CENSUS PROJECT

Our second step towards the URAPP was research on the indigenous peoples of Siberia and the Urals. The Mansi, Khanty and Nenets had been under Russian rule since the 17th century and they now claimed the right to use their lands. Dr. David Anderson headed this research project at the University of Aberdeen in 2005–2008. Its main part was locating the Polar census primary sources in the Russian archives and their transcription. While most of the Polar census manuscripts were lost, fragments including those on the Khanty, Mansi and Nenets survived in the GASO archive (Glavatskaya, 2011c; Glavatskaya & Borovik, 2013). The Soviet 1926–1927 Polar census was an extension of the 1926 all-union enumeration and the organizers introduced a unique system of ethno-demographic registration allowing detailed ethnographic descriptions of the indigenous households. Its value for research was pivotal, since this census took place on the eve of Soviet social modernization. The scanned and transcribed Polar census materials became the basis for several studies (Anderson, 2011).

The Polar census itself was an outstanding statistical and ethnographical study especially in the Obdor region far north in the Urals. The collected data permits understanding of the cultural landscape along ethnic, demographic, social, economic and religious dimensions. For the only time in Russia, the lives of most of the indigenous people were documented comprehensively. The Polar census covered a huge territory and collected the widest range of characteristics about people and settlements in the history of the census.

The census documents include a nominative household card with 405 cells, which lists each person by name (the household head also with a nickname), age, family relationship, marital status, ethnicity, occupation, income, etc. The census takers gave a qualitative description of the settlements in the settlement cards, providing detailed information on the settlement's exact location, its economy and involvement in trade. These cards also provide information about schooling and medical care, as well as traditional religion issues, shamans and healing practices. The budget cards were nominative and contain a thorough description of dwellings, the interiors, transportation means, number of the reindeer herds, pets, clothes, utensils, etc. The trade cards inform us about what the households sold, bought, exchanged and the quality of the game, specifying the different types of game and fish. It also includes a simple time use study, what equipment they needed, etc. (Glavatskaya, 2011c).

After the household cards, we transcribed the settlement cards, the budget cards and the trade cards for the Obdorsk region (Glavatskaya & Borovik, 2013). The polar census database became the basis for research on family patterns and polygamy in particular (Glavatskaya, 2015). To understand the details of the region's polygamy we needed the revizskie skazki, which were transcribed as part of the URAPP.

3.3 SERGEI SERGEL'S FIELD RESEARCH AMONG THE SAMI COMBINED WITH NORWEGIAN MICRODATA

We used an interesting publication by the Russian student-ethnographer Sergei Sergel, who spent several months together with the Norwegian Sami in 1907–1908. He joined the nomadic Sara family, accurately describing each family member, their relations, way of life etc. (Sergel, 1927). What happened with this family before and after Sergei Sergel met them, could be found in the transcribed Norwegian censuses and church records (Glavatskaya & Thorvaldsen, 2013). This research again demonstrated the importance of the nominative sources and their potential for studying ethno-religious minorities, also in Russia. That was the third main source of inspiration towards building the URAPP. The University of Aberdeen project focused on the Polar area of the USSR where the Kola Peninsula included a minority called Fil'mans. We shall return to this Sami minority in section 5.3 when we discuss marriage patterns.

4 THE URAL POPULATION PROJECT — CREATION, CONTENT AND STRATEGIES

4.1 CREATION

However, the creation of the URAPP cannot be understood without rewinding to 1998–1999, when Elena Glavatskaya was invited to the Norwegian Center for Advanced Study in Oslo to participate in the international project "The Endangered Language of Shamanhood" (Pentikäinen & Simoncsics, 2005). While in Oslo, she had a chance to be acquainted with demographers, who studied mortality (Hubbard et al., 2002). Naturally, historical demography could be combined with religious studies.

Shortage of documents on the history of religious groups and minorities led us towards the rich *metricheskie knigi* — the parish records. Having these data on religious groups in abundance, and the inspiration of the Western demographic databases such as IPUMS, CEDAR, and HSN, led to the ambitious idea to create URAPP — the Ural Population Project. The Ural Federal University supported our plan, and we received seed funding to establish the International Demographic Unit (IDUN), see <https://idun.urfu.ru/en/about-idun/>. The Russian Science Foundation enabled us to create two databases covering our research interests dealing with the Siberian indigenous peoples and the Ural religious landscape.

Initially, the local archive possessing the parish records did everything to block our access to the source material for scanning, but we overcame that obstacle. Another archive contained the *revizskie skazki* and eagerly provided us with them for a decent price. As a result, we made two separate databases. One based on the *revizskie skazki* collected in 1852 and 1858 in Obdorsk (contemporary Salekhard) region, the northernmost part of the Urals, among the Nenets and Khanty people — Siberian, indigenous reindeer herders. The other was based on parish records, covering the period 1900–1919. These two databases were the start of the URAPP which was extended with more data during the next stages.

4.2 CONTENT

The core of the URAPP are the transcribed parish records of Ekaterinburg, extended with records of Verkhoturie — an Orthodox Church center to the north of Ekaterinburg. In addition, there are the data from the village of Kislovskoe to the east where infant mortality (IMR) reached nearly 70% in the late 19th century, a figure based on the baptismal and burial lists transcribed into URAPP. There are also several data sets transcribed from Jewish parish records in Siberia and the Urals as well as other nominative sources, such as the synagogues' member rosters or lists of repressed Jews. Another part are the *revizskie skazki* of 1852 and 1858 from the Obdor region with indigenous population groups and a sample prepared for the MOSAIC project from the village of Sarana¹. Currently we are working on transcribing the 10th tax revision of 1856–1858 for Ekaterinburg. In addition, we found primary cards from the 1959 Soviet Union census for several districts in Sverdlovsk² and transcribed them. For an overview of all available data, see tables 3 to 7.

We primarily base our microdata research on vital events registered in the parish records (*metricheskie knigi*) from Ekaterinburg's religious communities. The ecclesiastical registration of vital events took place all over the Russian Empire until the October Revolution in 1917, with extensions where the white troops prevailed. After that, a civil office established by the Bolsheviks gradually took over the registration in each municipality.

We found parish records in the State Archive of Sverdlovsk oblast' (GASO) with births, weddings, funerals and divorces. The registration of marriages in the church books provides names (first, family and patronymics), marital status, social standing and/or occupation of grooms and brides (see section 5.3), their place of origin or registration, age, religion (when appropriate) and date of the wedding, information on the parents, witnesses and the person who performed the ritual.

The entries on burials in the parish registers provide names, death date, the age of the deceased and death cause. In the case of children aged under 16, there is also information about the parents: their names, social standing/occupation, place of birth or origin and marital status. In addition, the records contain data on priests conducting funeral services and occasionally death certificate extracts verified by a doctor or police officer.

Records on baptisms provide information on dates and places of birth and baptism, parents' names, their social status/occupation, place of origin and marital status and the same information on godparents. In addition, it included the names of the priests who conducted the baptism service and occasionally godparents' signatures. We have so far transcribed over 65,000 vital events for 10 Eastern Christian parishes of Ekaterinburg (Russian Orthodox Church, Edinovercyy and Old Believers), as well as the Catholic parish of St. Anna Church; St. Paul Lutheran Church; the Synagogue and the Muslim community (see Table 3).

- 1 <https://censusmosaic.demog.berkeley.edu/home>. However, Sarana data has not yet been processed and harmonized.
- 2 Ekaterinburg was renamed after the revolutionary leader Iakob Sverdlov in 1924. Only in 1992 its original name Ekaterinburg was restored.

Table 3 *Ekaterinburg parish records included in the URAPP database*

Parish	Denomination	Years	Baptisms	Weddings	Funerals	Divorces	Total
Ascension Church	Russian Orthodox	1880–1919	7,808	2,793	8,275	0	18,876
St. Epiphany Church	Russian Orthodox	1880–1919	5,976	1,100	4,273	0	11,349
St. Catherin Church	Russian Orthodox	1880–1919	1,874	2,613	7,386	0	11,873
Holy Spirit Church	Russian Orthodox	1880–1919	3,922	1,683	2,401	0	8,006
St. Alexander Nevskii Church	Russian Orthodox	1897–1919	1,463	365	3,505	0	5,333
Church of the Saviour	Edinovercty	1901–1919	497	278	515	0	1,290
Holy Trinity Church	Edinovercty	1901–1918	139	247	189	0	575
Assumption Chapel (Chasovennye)	Old Believers	1908–1919	185	41	125	0	351
St. Nikolai Chapel (Chasovennye)	Old Believers	1907–1919	246	24	194	0	464
Holy Trinity Church (Belokrinitskie)	Old Believers	1907–1926		129	347		476
St. Anna Church	Catholic	1898–1919	821	273	699	0	1,793
St. Peter Church	Lutheran	1886–1919	406	427	511	0	1,344
Synagogue	Jewish	1906–1917	520	136	219	18	893
Muslim Prayer house	Muslim	1891–1918	1,326	76	1,355	22	2,779
Total			25,183	10,185	29,994	40	65,402

Note: The parish records were transcribed from sources archived in the Gosudarstvennyi Arkhiv Sverdlovskoi Oblasti [State Archive of Sverdlovskaja oblast'], Ekaterinburg, Russia (GASO).

Apart from Ekaterinburg we have transcribed parish records of Verkhoturie — one of the oldest Russian cities in Perm' province, which used to be an administrative centre in 17th-century Siberia and the Orthodox Church parish with the biggest monastery and a pilgrimage site since the 18th century. Another sample of Orthodox Church parish records transcribed for URAPP are the records of St. Peter and Paul Church in Kislovskoe village, noted for its high infant mortality rate reaching 700‰ (see Table 4).

We also discovered synagogues' parish records in the Ural and Siberian regional archives and transcribed them (see Table 5). In addition to Ekaterinburg we transcribed revizskie skazki from Sarana settlement founded as metal producing factory in 1758 some 200 km to the west of Ekaterinburg and from the Obdorsk region in the polar area (see Table 6 and Figure 1 for location). Another sample of individual level data are the lists of Jewish settlers and members of synagogues, which were prepared by the authorities on different occasions and have a different structure. We transcribed these sources both for research purposes and in support of genealogical studies (see Table 7).

Table 4 *Verkhoturie (Intercession church) and Kislovskoe village (St. Peter and Paul church) parish records included in the URAPP database*

Parish	Denomination	Years	Baptism	Wedding	Funeral	Total
Intercession Church	Russian Orthodox	1886-1919	2,146	766	2,276	5,188 ³
St. Peter and Paul Church	Russian Orthodox	1880	158	46	166	370
St. Peter and Paul Church	Russian Orthodox	1915-1917	540	99	584	1,223
Total			2,844	911	3,026	6,781

Note: The parish records were transcribed from sources archived in the Gosudarstvennyi Arkhiv Sverdlovskoi Oblasti [State Archive of Sverdlovskaja oblast'], Ekaterinburg, Russia (GASO).

3 The transcription was done for the Mosaic project in collaboration with Max Plank Institute. Dr. Benjamin Matuzak while in Ekaterinburg as a research fellow in IDUN did significant parts of the transcriptions.

Table 5 *Urals and Siberian Jewish parish records included in the URAPP database*

City	Years	Birth	Marriage	Funeral	Divorce	Total	Archive
Troitsk	1906–1911	53			3	56	OGACHO ^a
Cheliabinsk	1877–1919	628	188	142	12	970	OGACHO ^a
Sarapul	1903–1917	105	10	7		122	TSGA UR ^b
Perm'	1880–1916	492	201	301	14	1,008	GAPK ^c
Ekaterinburg	1906–1917	520	139	18	219	896	GASO ^d
Tomsk	1860–1917	5,858	1,923	4,533	339	12,653	GATO ^e ; OGKU GATO ^f
Total		7,656	2,461	5,001	587	15,705	

Note: The parish records were transcribed from sources archived in the following places:

- ^a OGACHO — *Ob'edinennyi Gosudarstvennyi Arkhiv Cheliabinskoi Oblasti [United State Archive of Chelyabinsk Oblast'] in Cheliabinsk;*
- ^b TSGA UR — *Tsentral'nyi Gosudarstvennyi Arkhiv Udmurtskoi Respubliki [Central State Archive of Udmurt Republic] in Izhevsk;*
- ^c GAPK — *Gosudarstvennyi Arkhiv Permskogo Kraia [State Archive of Perm' Region] in Perm';*
- ^d GASO — *Gosudarstvennyi Arkhiv Sverdlovskoi Oblasti [State Archive of Sverdlovskaiia oblast'] in Ekaterinburg;*
- ^e GATO — *Gosudarstvennyi Arkhiv Tjumenskoi Oblasti [State Archive of Tjumen' Oblast'] in Tjumen';*
- ^f OGKU GATO — *Gosudarstvennyi Arkhiv Tomskoi Oblasti [State Archive of Tomsk Oblast'] in Tomsk.*

Table 6 *Data transcribed from revizskie skazki (census-like tax revisions) included in the URAPP database*

Place name	Year	Category	Men	Women	Both	Households	Archive
Ekaterinburg	1857–1858	Smal guild artisans [tsekhovye meschane] and merchants [kuptcy]	675	710	1385	256	GASO ^a
Sarana	1858	Peasants	1,435	1,416	2,852	487	GASO ^a
Obdorsk region Nenets and Khanty	1850–1852	Indigenous nomads	4,652	4,063	8,725	1,151	GATO T ^b
Obdorsk region Nenets and Khanty	1857–1858	Indigenous nomads	5,230	4,547	9,777	1,548	GATO T ^b
Total			11,992	10,736	22,739	3,442	

Note: The parish records were transcribed from sources archived in the following places:

- ^a GASO — *Gosudarstvennyi Arkhiv Sverdlovskoi Oblasti [State Archive of Sverdlovskaiia oblast'] in Ekaterinburg;*
- ^b GATO T — *Gosudarstvennyi Arkhiv Tjumenskoi Oblasti v Gorode Tobol'ske [Tobol'sk Branch of the State Archive of Tjumen' Oblast'] in Tobol'sk.*

The URAPP data set from the 1959 census for Ekaterinburg includes two parts: one containing personal data and another with aggregate data on the level of the family. The first one includes the following information: surname, first name and patronymic, age, sex, relationship to the householder, marital status, ethnicity (natsional'nost' in Russian), citizenship, mother tongue, educational level, place of work, position, source of subsistence, employer, occupation, and permanent place of residence. The second one builds on the first and contains number of family members, number of children under 18 in each family and number of employed family members (Gorbachev, 2020). So far, our 1959 transcription includes 9,382 persons (4,667 men, 4,715 women) in 2,079 families.

Table 7 *Lists of Jews in the URAPP database*

Document	Years	Entries	Archive
Tjumen' city Jews	1907, 1911–1912	50	GATO ^e
Ekaterinburg city Jews	1901	145	GASO ^d
Perm' province Jews	1901	172	GAPK ^c
Perm' city Jews	1901	256	GAPK ^c
Ekaterinburg city Jewish communities leaders	1920	36	GASO ^d
Ekaterinburg city Jewish communities leaders	1925	131	GASO ^d
Ekaterinburg city Jews	1927	374	GASO ^d
Repressed Jews in Sverdlovskaja oblast'	1917–1938	423	GASO ^d
Total		1,587	

Note: For the archives where the parish records were transcribed from, see the explanation at table 5.

4.3 STRATEGY

The strategy of the URAPP project has evolved from one of transcribing microdata about religious minorities to one of covering the whole population. The research experience of the group was to work with the Orthodox minority of Old Believers and ethno-religious minorities, as well as indigenous peoples in the northern Urals. It was then natural to seek funding for transcribing church records covering areas inhabited by these minorities, in order to complement previous findings with new results based on the vital records. This led to a comparative deficit, and we saw the need for including microdata about the Orthodox majority from their parish records and the revizskie skazki tax censuses — the manuscripts of most modern censuses are not available. Therefore, we included several parishes from Ekaterinburg city, Verkhoturie town — the center of Orthodoxy in the Urals, as well as a village of Kislovskoe. In the long run, we would like to extend the URAPP into a nationwide register, however right now we are planning to extend the data entry in the city longitudinally and to include a smaller factory town in the Urals to run comparative analyses of infant mortality.

The project on the Jews in the Urals and Siberia is a PhD student project. In addition to the research, it has the rather ambitious aim to eventually construct a dataset on the understudied Jewish population who lived beyond the Pale of Settlement, in the Urals and Siberia. In order to find more resources, we are negotiating with genealogists, but targeting research council funding in connection with research projects will still be important. We clearly see a need to expand the database in both time and space, especially to cover the 1920s with data from the civilian life event registers.

Another priority is to cover wider areas of Ekaterinburg and the surroundings. Methodologically, we have started experiments with record linkage. Another task is to systematize the encoding performed in connection with concrete research tasks into standardized translation tables. We have also built a web-based user interface, where genealogists can trace their ancestries with an exemption for the 1959 census data which are too recent to be made publicly available.

5 RESEARCH WITH THE URAPP DATABASE

Our initial research interests focused on religious and ethnic minorities and directed the development of the databases and the URAPP in general. No wonder most of the research conducted by the URAPP team members has ethno-religious issues as its main topic, combined with research questions on migration, nuptiality and mortality issues. Below are examples of research based on the URAPP data.

5.1 JEWISH STUDIES

Our study of Jewish history in Ekaterinburg is based on the nominative vital records of the Orenburg battalion No. 8 and the Synagogue's parish records (see Table 3). These documents registered the Jewish marriages from the first families of Jewish soldiers in Ekaterinburg in 1850. According to our

data, Ekaterinburg's Jews managed to keep connections not only with relatives remaining within the Pale of Settlement in the western provinces of the Empire, but also with Jewish communities in western Siberia. With few exceptions, they created ethnically and religiously homogeneous marriages, which contributed to the preservation of their ethnic and religious identity. Observing religious regulations with regard to the time and date of marriage at least until 1917, each marriage was accompanied by the signing of a marriage contract, the so-called *ktuba*. The presence of a government rabbi was not mandatory; instead, the so-called spiritual rabbis or respected members of the community could conduct the wedding. Despite the pressure of the authorities against religion during the Soviet era, some Sverdlovsk Jews continued to hold religious weddings with the *ktuba* signing, putting up the wedding canopy *huppah* and breaking glass in memory of the destroyed Temple in Jerusalem. After 1917, they registered marriages and divorces in the secular state's vital events registry offices (Glavatskaya & Zabolotnykh, 2018). Jews' conversion to Russian Orthodoxy in Ekaterinburg in the early 20th century was also studied. The highest number of such Jewish baptisms was recorded in 1911–1912. The decision to accept Orthodoxy was taken familywise, including infants. However, most often this decision was made by young people in their twenties since religious affiliation was part of the wedding preparations or career plans (Zabolotnykh, 2018).

5.2 THE OLD BELIEVERS STUDIES

The Old Believers are ultra-conservative dissenters who split from the Russian Orthodox Church in the 17th century. Persecuted by the state, they migrated to the country's peripheries, including the Urals, in order to maintain their pre-reform traditions. This biggest religious minority among ethnic Russians got legal status and started to register vital events in church books after the Religious Freedom Manifesto of 1905. Complementing an extensive bibliography on the Urals Old Believers history and culture, the URAPP allowed us to conduct several studies on their demography. We have presented results from the computerized analyses of their parish records, including marriage activities, age at first marriage with special attention to gender, social status and migration as determinants of marriage timing. We also addressed the issue of remarriage and conversion in connection with marriage, and argue that it was a sign of social lifting and abandoning of religious endogamy, signaling modernization of the marriage institution in early 20th-century Russia (Borovik, 2018, 2019b; Glavatskaya & Borovik, 2019; Palkin & Borovik, 2019).

With the birth entries, Julia Borovik analyzed the Old Believers' practice of naming their newborn and found evidence of modernization. Ekaterinburg's Old-Believers more often followed the current naming fashion, in addition to the Russian Orthodox Church prescription to use the name of the Saint, whose veneration day was closest to the baby's birth or baptism (Borovik, 2019a). In a parallel study on the Orthodox, Elizaveta Zabolotnykh disclosed an interesting practice of giving identical names to both twins and to all three children in a triplet (2020).

5.3 NUPTIALITY STUDIES

5.3.1 RELIGION AND AGE AT FIRST MARRIAGE

Our analyses of marriage behavior in the church records allowed insights into different religious groups' life in early 20th-century Ekaterinburg. The individual level provided clues to understand the religious affiliations' influence on age at first marriage. The Catholic, Lutheran and Muslim men married a few years older than the Russian Orthodox and Jewish men. The difference in age at first marriage of their brides was less significant: one to two years higher than Catholic and Lutheran women compared to the Russian Orthodox and one year younger than the Muslim brides. We also found that belonging to a certain parish within the same confession might influence mean age at first marriage in case of both brides and grooms (Glavatskaya, Borovik, & Bobitskii, 2016; Korkodinova, Glavatskaya, & Borovik, 2016).

Unfortunately, we do not yet have individual level data after 1919 to explain why 10 years after the Revolution Sverdlovsk abandoned the European tradition of late marriage. According to the 1926 aggregate census data, both the singulate mean age at marriage (SMAM) and the percentage of never married declined. Our hypothesis is that after the Revolution, the new regime shut down in-migration from the West, while in-migration increased from the rural suburbs. These in-migrants, paradoxically during revolutionary times, brought with them Orthodox or Muslim marital ethics, which required obligatory marriage. The Bolshevik legislation separated the Church from the State, depriving it of the right to register weddings, and introduced freedom of divorce in 1917. Both marriages and divorces

were to be registered by a civil officer in a state office, with no sacraments required. To prove this hypothesis we need individual level data from the new civilian sources, which are part of our next project (Glavatskaya, Bobitsky, Zabolotnykh, & Vishnevskaya, 2019).

5.3.2 RELIGION, MARRIAGE AND WORLD WAR I

We also conducted research on prisoners of war (POWs) who were kept in the Urals from 1915 to 1919: their numbers, their nationality and their marriage strategies. During World War I, the Urals received both refugees and prisoners of war, many of whom were Catholics, Lutherans and Jews. Individual level data show that mean age at first marriage was slowly increasing in the State Church parishes until World War I. With the war, the average age at first marriage increased by 1.8 years for grooms and 1.3 for brides. The Lutherans, mostly ethnic Germans, postponed their weddings even more, by 2.3 and 2 years respectively; likely due to the shortage of eligible partners on the marriage market since the Lutheran-German population decreased dramatically between 1913 and 1920. Naturally, their life in a country being at war with Germany was difficult. We found that the POW groups joined the marriage market of Ekaterinburg from 1916, influencing the city's demography to varying degrees, and that religious affiliation played an important role for the demographic consequences and intermarriage (Glavatskaya & Borovik, 2016; Glavatskaya, Borovik, Thorvaldsen, & Zabolotnykh, 2020).

5.3.3 MIXED MARRIAGE

Our comparative analyses of mixed marriages with respect to religion and ethnicity focused on Russia and Norway. In both countries, the State Churches dominated religious life with more than 90% of the population during the decades around 1900. However, both were losing influence during this period — rapidly in Russia after the 1917 revolution. The Finno-Ugric Sami and Finns were the main ethnic minorities in Norway, while Russia also had over 100 other ethnic groups. The research on Norway employs nominative and aggregate census material, which from 1865 asked questions about religious affiliation, while the Russian case study utilized the database of church microdata being built for Ekaterinburg, in addition to census aggregates. Our main conclusion is that religion was a stronger regulator of intermarriage than ethnicity. Thus, the Lutheran Sami on the Kola Peninsula, who never intermarried with the Orthodox Sami according to the 1926 Polar census microdata, were typical. Religious intermarriage was also unusual in Ekaterinburg, even if official regulations were softened by the State over time. The exception was during World War I, when there was a deficit of young, Russian men at home and an influx of refugees and Austro-Hungarian Prisoners of War (mostly Catholics and Lutherans). In addition, the 1917 Revolution created equal rights for all religious denominations. The relatively few religious intermarriages in Norway were mostly between members of different Lutheran congregations — atheist men being the only group who often outmarried (Glavatskaya, Thorvaldsen, Borovik, & Zabolotnykh, 2020).

5.3.4 INDIGENOUS POLYGAMY

A particularly interesting nuptiality study was performed on the Khanty and Nenets — indigenous groups in northwestern Siberia based on the URAPP dataset with entries from the 9th revision of 1852 and the 10th revision of 1858. This transcribed dataset contains 1,240 households, 450 Khanty and 790 Nenets. Since there were 4,280 men and only 3,771 women registered, we assume significant female under-registration. According to the database, there were 1,604 married men and 1,712 married women, which is not surprising, since there were 105 cases of polygyny reported: 20 among the Khanty and 85 among the Nenets. As usual, polygyny ran in certain families: in most cases, if the head of the household had more than one wife, so did his brothers and sons, if living together. The maximum number of wives — three for one man, was recorded in seven households. According to our data, up to 7% of the married men had more than one wife in 1858. Thus, in the middle of the 19th century polygyny was officially recognized and a common phenomenon among the Khanty and especially among the Nenets of the Obdorsk region (Glavatskaya, 2015).

5.4 RELIGION AND THE URBAN FAMILY

In Russia, the rural family has been studied in detail with ethnographic methods, but the urban family previously received little attention. Ekaterinburg, chosen as the object of study, in the mid-19th century became the largest industrial center in Perm' province. We used the materials of the 10th revizskie skazkie of 1858. The households were divided into three categories, reflecting their socio-economic status and formal social standing (so-called *soslovie* [estate]) in the city. The first category

kuptcy [merchants], included both owners of trading enterprises and owners of small shops. Like the nobility, they were exempt from taxation, but unlike the nobility still registered. The second category, *meschane*, were petty bourgeois and former peasants, who became city dwellers engaged in trade and services both as office clerks and workers. The third category was the *tcekhovye meschane* [guild craftsmen], engaged in craftsmanship.

Analysis of the tax census data from 1858 showed that 39.7% of the households with 28.4% of the inhabitants were nuclear families. Almost the same proportion of households, 38.9%, were family households that included more than two generations of relatives or several married couples related by family ties (19.7% and 19.2%, respectively). In these extended households lived 25% and 39.6% respectively of the city dwellers. A less significant number of households, 15.3%, consisted of just one person whose other family members still lived in the countryside. The average size of Ekaterinburg households was five persons. With respect to social standing (“estate”), the petty bourgeois, with business activities based on family cooperation, averaged 6.2 persons, and the merchants averaged 4.2 persons. Among the guild craftsmen, combined group and family households in the city included up to 25 persons. Our analyses allowed us to determine the religious composition among the representatives of this social and professional group and their family size. The high share of Old Believers among the guild craftsmen was related to this confession by ancestry, they continued to live according to the rules of this religious community that had developed in the 18th and 19th centuries (Borovik & Glavatskaya, 2020). This study confirmed both the importance of the religious factor and its influence on preserving the large family tradition (Szołtysek, 2015). However, we also argue that big households were justified by the necessities of production.

5.5 MORTALITY

The mortality situation in late 19th to early 20th centuries Ekaterinburg was influenced by the mass influx of industrial and craftsman enterprises as well as migrants, which worsened the already poor environmental and sanitary conditions. The city administration was unable to provide urban dwellers with health care, access to information and sometimes even food, which resulted, among other things, in a high level of infant mortality rates. Thus, Ekaterinburg was lagging more and more behind the regional standards at the time. The role of ethnic and religious affiliation was a significant factor as it determined either a high level of education (Jews, Catholics, Lutherans) and/or strict adherence to hygiene rules (Jews, Muslims). These factors, in their turn, determined the quality of infant care and increased the baby's chances to survive in the city. Peasants who constituted the majority in Orthodox parishes brought potentially hazardous rural ways of life into the densely populated city and had lower levels of education. They were suspicious of doctors, preferring folk healing practices or simply had no time or money to obtain professional medical assistance.

5.5.1 INFANT AND CHILD MORTALITY

Demographers analyze regional and other infant mortality differentials as important factors behind the current life expectancy of Russian citizens (Kumo, 2017). Historically, however, the Russian Empire is simply displayed as one block with high infant mortality rates (Kluesener et al., 2014). The first epidemiological transition started in Russia later than in most European countries and soon after the start was interrupted by the socio-political disasters of the early 20th century (Isupov, 2016). Although the administrative region of Perm', surrounding Ekaterinburg, had extremely high levels of infant mortality when it entered the epidemiological transition, it soon became one of the leaders in terms of declining infant mortality rates (IMR). While from 1886–1897 to 1908–1910 IMR declined on average at a pace of 21‰ points across Russia, the Perm' gubernia IMR dropped by 117‰ points. Comparative analysis of district and city dynamics shows that the IMR declined in rural areas while in the city it remained on the same level.

We believe that this effect was due to the doctors that were employed by the *zemstvo* (self-governing, elected, sub-provincial level institutions introduced in 1865 to manage local affairs including medical services, sanitation, public education and other socially important activities). These doctors focused predominantly on promoting knowledge and medical care in rural areas. This movement was particularly influential in the Urals, which had a large number of *zavody* (metal producing factories) with a surrounding population which was generally more exposed to innovations. These settlements had a developed medical network, a system of district doctors (Shestova, 2017, p. 38) and nurseries (Golikova & Dashkevich, 2014). Spatial analyses of the IMR in Ekaterinburg *uezd* [counties] supports

this hypothesis. The uezds' subdivisions with industrial settlements on their territories had lower IMR than its agricultural units (Bakharev, 2017).

Our hypothesis that the level of infant mortality is closely connected with the type of settlement was confirmed when we compared the corresponding data for a city parish and the surrounding countryside. Late 19th-century Ekaterinburg had a moderate level of infant mortality, but from 1889 to 1917 it demonstrated only a slight decrease in the post-neonatal infant mortality rate. This decrease, however, was nullified by a slight increase in neonatal mortality: 1909, 1911 and the first year of war — 1914 — were grievous years for infants, which affected IMR (see Figure 5).

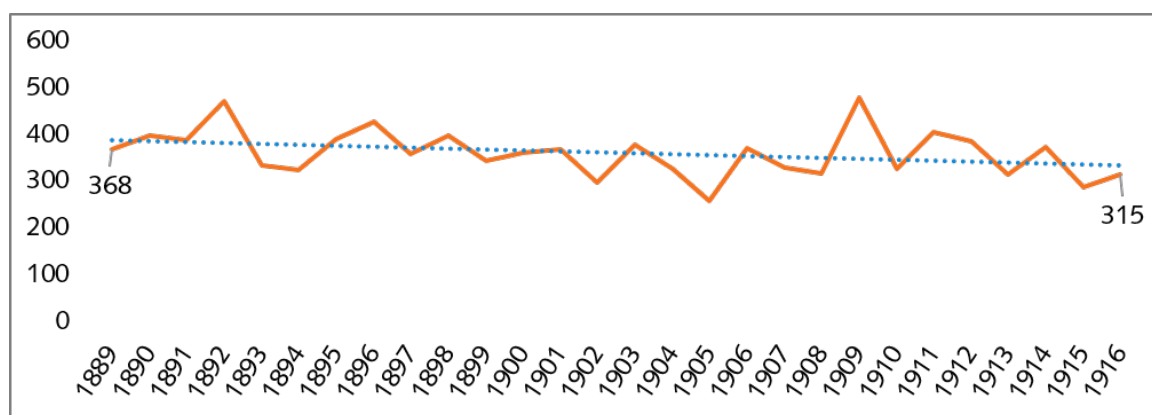
Our analyses of IMR among the different religious communities in Ekaterinburg showed that religion did matter: Jews and Catholics, minorities with higher education and cultural background from the western parts of the Empire, did better with respect to infant mortality around the start of the 20th century. The Orthodox minority of Old Believers was in a middle position, but clearly did better than the Orthodox majority (Glavatskaya, Borovik, & Thorvaldsen, 2018).

5.5.2 CAUSES OF DEATH

We were able to identify the main causes of child mortality from the parish records from two Orthodox parishes in Ekaterinburg — Ascension and Epiphany — for the period 1880–1919, comprising 7,187 records. We encoded the records in accordance with the international historical death causes classification, developed by European historical demographers (Sommerseth & Walhout, 2019). It consists of six classes: deaths caused by infections (1); non-communicable diseases (2); causes described by outdated popular terms (3); external causes such as drowning (4); illegible records (5); missing causes (6). Our study revealed low involvement of professional physicians in the mortality registration — doctors checked less than 4% of childhood deaths. In Ekaterinburg, the clergy performed this duty, and an indication of the specific cause of death was stated in almost all records. Analysis of the Ekaterinburg data showed that infectious diseases claimed up to 65% of children's deaths; diseases described with obsolete terms caused 28% of children's deaths and non-communicable diseases claimed 7% of the deaths among young Ekaterinburg residents. Deaths due to external causes amounted to less than 1% (Bakharev & Glavatskaya, 2019).

The URAPP data in addition allowed us to find out that the members of the Ascension church parish on average baptized their babies within the first three days of life (see Table 8).

Figure 5 *IMR Dynamics per 1000 live births in the Ascension Church parish, 1889–1916**



* Instead of the missing 1904 data, we inserted aggregate numbers.

Source: Bakharev & Glavatskaya (2019)

Table 8 *Average baptism age of infants in the Ascension Church parish, Ekaterinburg*

	1889–1899	1900–1910	1911–1918
Average baptism age in days	3,7	4,6	5,2

Source: Bakharev & Glavatskaya (2019)

This practice was based on the Orthodox Church's strong belief that babies would not reach paradise if they died before being baptized. This belief made both parents and priests hurry with baptism, whether a baby was well or not. Given the fact that the ritual required a baby's complete immersion into a vessel full of water three times, it is easy to believe that the whole procedure could affect the babies' health. It is interesting to note, that Ekaterinburg Catholics, who generally believed in the same idea that only baptized babies would get access to heaven, on average baptized their newborn at the age of 41 days (Bakharev & Glavatskaya, 2019). In accordance with other researchers we found that the especially high infant mortality was also caused by inconsistent breastfeeding and lack of hygienic measures for childcare (Ransel, 1991).

In order to study the non-specific cause of death "old age", we split all deaths into three groups: Russian Orthodox Church parish members, religious minorities (migrants from European Russia: Catholics, Lutherans and Jews) and Muslims (local migrants from rural areas). The analysis showed that Russians aged over 50 were registered as dead from "old age" more often than the minorities from the west and local Muslims. In all three groups, women more often received the cause of death "old age" than men, and the widest gender difference was among Muslims. We interpret this as caused by the European origin of the Catholic, Lutheran and Jewish migrants, who were more educated and more likely to seek medical help when necessary.

5.5.3 THE SPANISH FLU IN EKATERINBURG

Inspired by the COVID-19 pandemic, we focused on the Spanish flu pandemic of 1918–1920, which killed, according to some researchers, from 50 to 100 million people while others estimate lower numbers such as 20 million victims (Johnson & Mueller, 2002). The reason for this uncertainty is that data for a number of countries, including Russia, are rather rough estimates based on mortality rates from other parts of the world. We analyzed the causes of death in Ekaterinburg during the period of the Spanish pandemic to determine likely signs of the spread of influenza (Glavatskaya & Thorvaldsen, 2020). The church records indicate that the Spanish flu affected mortality levels in Ekaterinburg during and after the end of World War I. This applies particularly to the timing of the seasonal mortality spikes and to the many who died from respiratory illnesses. However, these features dominated only to some degree, and they did not create such clear and special gender and age profiles as in Western Europe and the US. This is likely because most young men were away fighting in the army, people were protected by the enormous distances and because train traffic was disrupted due to the hostilities

Another hypothesis is that persons aged over 30 had immunity due to the "Russian flu" pandemic in the winter of 1889–1890, which may explain why younger men were more at risk in 1918 (Shanks & Brundage, 2012). We lack historical mortality statistics covering larger areas, but further microdata from parish records can be brought forward from the archives to illuminate the course of the Spanish influenza in Russia. Based on our present evidence, we can only conclude that the Spanish flu virus in all likelihood hit Russia with less force than the US or Norway at 0.6% mortality, and we agree with Patterson and Pyle (1991) that it killed less than half a million persons. While global mortality has been estimated at 2.5% (Billings, 1997), this gives a mortality rate under a half percent in a Russian population of 137 million.

6 CONCLUSION

We are building the Ural Population Project (URAPP) from individual level data transcriptions of the 19th- to early 20th-century parish records (metricheskie knigi) and the mid-19th-century census-like tax revisions manuscripts. Decade-long studies of the ethno-religious cultural landscape of the Urals and northwestern Siberia with qualitative methods and sources are an important background of the URAPP. The traditional quantitative focus on statistical aggregates in Russia contributed little to our understanding of this research field. However, contacts with and participation in demography and social history projects abroad and in Russia, showed the potential of using individual level data combined with quantitative and qualitative methods. The microdata could be combined in flexible ways, aggregated to the unit level studied and even be used to illuminate groups of people who had been studied in previous research. Of special significance on our way towards the URAPP were the transcriptions and studies of the 1926–1927 Polar census manuscripts, the world's most detailed

census ever, focusing on the ethnic minorities in the northern parts of Russia. The Polar census does not cover the Sverdlovsk region, however, and has not become a priority for the URAPP.

The more than 65,000 vital events transcribed from parishes of Russian Orthodox Churches and minority religions in and around Ekaterinburg have been the basis for studies of mortality and nuptiality, in combination with the parishioners' religion and other characteristics. We found that the Jewish population in Ekaterinburg managed to hold on to their traditions and to keep connections both with relatives in the western Pale of Settlement as well as further east and that they usually married within their ethnic group. Catholic, Lutheran and Muslim men married a few years older than the Russian Orthodox and Jewish men did, and both the levels and the difference increased during World War I. Prisoners of war joined the difficult marriage market of Ekaterinburg from 1916, usually marrying within their own religious group. Also, the comparison of marriage strategies in Russia and Norway proved religion to be more decisive than ethnicity when finding a marriage partner. Based on the census-like tax revisions from the mid-19th century, polygyny was common and officially recognized among the indigenous Khanty and Nenets of the Obdorsk region, as a rule running in certain families.

Analyses of infant mortality in the religious communities' parish records in Ekaterinburg showed that religion did matter: Jews and the Catholics, minorities with higher education and background from the west, experienced lower infant mortality around the start of the 20th century. The Orthodox minority of Old Believers was in a middle position but did better than the Orthodox majority. In addition to inadequate care and nutrition, an important reason was that the latter brought their newborn to church for an extremely early and tough baptism. The causes of death stated in the protocols also indicate that this religious majority less often sought help from medical doctors, more often simply stating "old age". Lastly, explorative analysis of the burial records for Ekaterinburg shows cases of the Spanish flu in 1918–1919, but on a lower level than what is found in the US or Western Europe, which supports recent theories that Russia was less hit by this pandemic than earlier non-empirical calculations indicated, and that some global estimates of flu mortality may be too high.

The strategy of the URAPP project has evolved from one of transcribing microdata about religious minorities to one of covering the whole population. We clearly see a need to expand the database in both time and space, and in this respect both genealogical, infrastructure oriented and research project resources must be combined. We plan to extend the URAPP by also transcribing the rest of the revision lists of 1858 for Ekaterinburg and pioneering the selected individual level vital events records from the 1920s. Further development of the URAPP and record linkage will allow us to base our findings on more vital records and cross-sectional data in order to also answer research questions on fertility and naming traditions.

ACKNOWLEDGEMENTS

The research was supported by the Russian Foundation for Basic Research (project number 19–29–07154).

REFERENCES

- Andreev, A., & Andreev, M. (n.d.). *Gde iskat' revizskie skazki po perepisi naseleniia* [Where to search tax revisions as population censuses]. http://livemem.ru/articles/revizskie_skazki.html
- Anderson, D. G. (Ed.) (2011). *The 1926/27 Soviet polar census expeditions*. Oxford, New York: Berghahn.
- Bakharev, D., & Glavatskaya, E. (2019). Infant mortality in the late 19th and early 20th century Urals: Macro and micro analyses. In E. Glavatskaya, G. Thorvaldsen, G. Fertig, & M. Szoltysek (Eds.), *Nominative data in demographic research in the East and the West* (pp. 202–219). Ekaterinburg: Ural University Press. doi: [10.15826/B978-5-7996-2656-3.12](https://doi.org/10.15826/B978-5-7996-2656-3.12)
- Bakharev, D. S. (2017). Mladencheskaya smertnost' v Ekaterinburgskom uezde v kontse XIX veka: Opyt kartografii [Infant mortality in Yekaterinburg uezd in the late XIX century: Experience of mapping]. Paper presented at the *International Scientific Conference "Digital Humanities: Resources, Methods, and Research"*. Perm', Russia. Retrieved from http://2017.dhconf.ru/wp-content/uploads/2017/05/DH_PERM_2.pdf

- Billings, M. (1997). *The influenza pandemic of 1918*. Retrieved from <https://virus.stanford.edu/uda/>
- Blum, A., & Troitskaya, I. (1997). Mortality in Russia during the 18th and 19th centuries: Local assessments based on the Revizii. *Population: An English Selection*, 9, 123–146. Retrieved from <https://www.jstor.org/stable/2953828>
- Borodkin, L. I., & Vladimirov, V. H. (2017). Asociaciija «Istorija i Komp'juter»: 25 let spustja [Association "History and Computing": 25 years on]. *Historical informatics*, 3, 1–6. doi: [10.7256/2585-7797.2017.3.24702](https://doi.org/10.7256/2585-7797.2017.3.24702)
- Borovik, I. V. (2018). Staroobriadtsy-chasovennye Ekaterinburga: Chislennost soslovnaia prinadlezhnost i proiavlennie konfessionalnoi obosoblenosti [The Old Believers of Yekaterinburg: Number, social status, and religious identity]. *Izvestiia Uralskogo federalnogo universiteta. Serii 2. Gumanitarnye nauki*, 20(1), 160–180. doi: [10.15826/izv2.2018.20.1.013](https://doi.org/10.15826/izv2.2018.20.1.013)
- Borovik, I. (2019a). Lichnye imena novorozhdennykh v ekaterinburgskikh staroobryadcheskikh obshchinah nachala XX veka [Personal names of newborns in the Old Believer communities of Ekaterinburg in the early 20th century]. *Voprosy Onomastiki*, 16(3), 30–47. doi: [10.15826/vopr_onom.2019.16.3.029](https://doi.org/10.15826/vopr_onom.2019.16.3.029)
- Borovik, I. (2019b). *Staroobryadcheskaya obschina i sem'ya rossiiskogo goroda: Ekaterinburg* [Old Believers' congregation and family life in the Russian town Ekaterinburg]. Ekaterinburg: Ural University Press.
- Borovik, I., & Glavatskaya, E. (2020). Tsekhovye meshchane Ekaterinburga po materialam X revizii: Religii i razmer sem'i [Guild craftsmen of Yekaterinburg according to 10th census: Religion and family size]. In *Dokumental'noe nasledie i istoricheskaja nauka. Materialy Ural'skogo istoriko-arhivnogo foruma, posvjashhennogo 50-letiju istoriko-arhivnoj special'nosti v Ural'skom universitete* (pp. 143–147). Ekaterinburg: Ural University Press. Retrieved from https://elar.ufu.ru/bitstream/10995/92827/1/978-5-7996-3078-2_2020.pdf
- Bryukhanova, E. A. (2019). Perepis' 1897 g.: Obretenie «utrachennykh» materialov i ikh predvaritel'nyi analiz [The 1897 census: The acquisition of "lost" materials and their preliminary analysis]. *Izvestiia Uralskogo federalnogo universiteta. Serii 2. Gumanitarnye nauki*, 21(3), 152–167. doi: [10.15826/izv2.2019.21.3.053](https://doi.org/10.15826/izv2.2019.21.3.053)
- Clem, R. S. (Ed.). (1986). *Research guide to the Russian and Soviet censuses*. Ithaca, NY: Cornell University Press. Available from <https://www.cornellpress.cornell.edu/book/9781501707070/research-guide-to-the-russian-and-soviet-censuses/#bookTabs=4>
- D'iachkov, V. L., Kanishchev, V. V., & Orlova, V. D. (2007). Mesto metriceskikh knig v komplekse istochnikov po istoricheskoi demografii Rossii XVIII – nachala XX v. [The parish registers in the complex of sources on the historical demography of Russia in the 18th – early 20th centuries]. In V. N. Vladimirov (Ed.), *Materialy tserkovno-prikhodskogo ucheta naseleniia kak istoriko-demograficheskii istochnik* (pp. 48–84). Barnaul: Altai State University Press.
- Glavatskaya, E. (1995). Christianization=Russification? On preserving the religious and ethnic identity of the Ob-Ugrians. In J. Pentikäinen (Ed.), *Shamanism and Northern ecology* (pp. 373–386). Berlin-New York: Mouton de Gruyter. doi: [10.1515/9783110811674.373](https://doi.org/10.1515/9783110811674.373)
- Glavatskaya, E. (2011a). The Mansi sacred landscape in long-term historical perspective. In P. Jordan (Ed.), *Landscape and culture in Northern Eurasia* (pp. 235–257). Walnut Cree, CA: Left Coast Press.
- Glavatskaya, E. (2011b). Siberian indigenous religious traditions in an ever changing world: The Khanty and Nenets case. In T. Yamada & T. Irimoto (Eds.), *Continuity, symbiosis, and the mind in traditional cultures of modern societies* (pp. 95–107). Sapporo: Hokkaido University Press.
- Glavatskaya, E. (2011c). Undaunted courage: The Polar census in the Obdor region. In D. Anderson (Ed.), *The 1926/27 Soviet polar census expeditions* (pp. 97–116). Oxford, New York: Berghahn.
- Glavatskaya, E. (2015). Polygamy among indigenous people of northern West Siberia in ethnographic and early census materials. *The History of the Family*, 21(1), 87–100. doi: [10.1080/1081602X.2015.1046467](https://doi.org/10.1080/1081602X.2015.1046467)
- Glavatskaya, E., Bobitsky, A., Zabolotnykh, E., & Vishnevskaya, A. (2019). Religion and marriage age in early twentieth century Ekaterinburg, Russia: A microdata analysis. In E. Glavatskaya, G. Thorvaldsen, G. Fertig, & M. Szoltysek (Eds.), *Nominative data in demographic research in the East and the West* (pp. 138–155). Ekaterinburg: Ural University Press. doi: [10.15826/B978-5-7996-2656-3.08](https://doi.org/10.15826/B978-5-7996-2656-3.08)
- Glavatskaya, E., & Borovik, I. (Eds.). (2013). *Ural'skaia ekspeditsiia na Obdorskom Severe: Pripoliarnaia perepis', 1926–1927 gg.* [Ural Expedition in the Obdorsk North: Subpolar census, 1926–1927]. Ekaterinburg: Ural University Press

- Glavatskaya, E., & Borovik, I. (2016). Death and marriage: World War I Catholic prisoners in the Urals. *Transylvanian Review*, 25(4), 28–40.
- Glavatskaya, E., & Borovik, J. (2019). The Old Believers and their marriage in the early twentieth century Urals, Russia: A microdata analysis. *Transylvanian Review*, 28(1), 112–130.
- Glavatskaya, E., Borovik, J., & Thorvaldsen, G. (2018). Urban infant mortality and religion at the end of the nineteenth and in the early twentieth century: The case of Ekaterinburg, Russia. *The History of the Family*, 23(1), 135–153. doi: [10.1080/1081602X.2017.1341845](https://doi.org/10.1080/1081602X.2017.1341845)
- Glavatskaya, E., Borovik, J., Thorvaldsen, G., & Zabolotnykh, E. (2020). From war to wedding: Marriage strategies of WWI POWs in the Urals, Russia. In S. Brée & S. Hin (Eds.), *The impact of World War I on marriages, divorces, and gender relations in Europe* (pp. 252–276). Leiden: Routledge. doi: [10.4324/9780429243684](https://doi.org/10.4324/9780429243684)
- Glavatskaya, E., & Thorvaldsen, G. (2013). Sergej Sergel's field research in Northern Norway and Finland: Contextualizing early 20th-century Sami. *Arctic Anthropology*, 50(1), 105–119. doi: [10.3368/aa.50.1.105](https://doi.org/10.3368/aa.50.1.105)
- Glavatskaya, E., & Thorvaldsen, G. (2015). Sibirskij Vavilon: Shvedskie uzniki v nachale XVIII v. *Quaestio Rossica*, 4, 215–240. doi: [10.15826/qr.2015.4.134](https://doi.org/10.15826/qr.2015.4.134)
- Glavatskaya, E., & Thorvaldsen, G. (2020). What role did the Spanish flu play? Analysis of the death causes in Ekaterinburg 1918–1919. In I. M. Garskova (Ed.), *Istoricheskie issledovanija v kontekste nauki o dannyh: Informacionnye resursy, analiticheskie metody i cifrovye tehnologii. Materialy mezhdunarodnoj konferencii* (pp. 33–39). Paper presented at the Association for History and Computing, Moscow State University. doi: [10.29003/m1786.978-5-317-06529-4/33-39](https://doi.org/10.29003/m1786.978-5-317-06529-4/33-39)
- Glavatskaya, E., Thorvaldsen, G., Borovik, I., & Zabolotnykh, E. (2020). Mixed marriages in late nineteenth to early twentieth century: Comparing Russia and Norway. *Journal of Family History*, 46(4), 414–432. doi: [10.1177/0363199020945215](https://doi.org/10.1177/0363199020945215)
- Glavatskaya, E. M., & Zabolotnykh, E. A. (2018). «...Po zakonu Moiseia i Izrailia»: Brak za chertoj osedlosti (po materialam evreiskoi religioznoj obshchiny Ekaterinburga) [Jewish Marriages outside the pale of settlement (with reference to the materials of the Yekaterinburg Jewish religious community)]. *Izvestiia Uralskogo federalnogo universiteta. Seriya 2. Gumanitarnye nauki*, 20(4), 68–84. doi: <https://doi.org/10.15826/izv2.2018.20.4.063>
- Glavatskaya, E. M., Borovik, J. V., & Bobitsky, A. V. (2016). Katoliki Ekaterinburga v konce XIX – nachale XX v. po materialam perepisej i metriceskikh knig [The Catholic community of Yekaterinburg between the late 19th and early 20th centuries according to the 1897 census and church records]. *Izvestiia Uralskogo federalnogo universiteta. Seriya 2. Gumanitarnye nauki*, 18(3), 68–84. doi: [10.15826/izv2.2016.18.3.044](https://doi.org/10.15826/izv2.2016.18.3.044)
- Golikova, S. V., & Dashkevich, L. A. (2014). Spasenie zhizni detey: Opyt ural'skikh guberniy v kontse XIX – nachale XX veka [Saving children's lives: The Ural provinces' experience in the late XIX – early XX centuries]. *Vestnik Permskogo Universiteta. Seriya Istorija*, 24(1), 124–134.
- Gorbachev, O. (2020). Vsesoiuznaia perepis' naseleniia 1959 g. kak istochnik dlia izuchenii istorii gorodskoi sem'i [USSR population census 1959 as a source for studying the history of the urban family]. In *Dokumental'noe nasledie i istoricheskaja nauka. Materialy Ural'skogo istoriko-arhivnogo foruma, posvjashhennogo 50-letiju istoriko-arhivnoj special'nosti v Ural'skom universitete* (pp. 154–158). Ekaterinburg: Ural University Press. Retrieved from https://elar.urfu.ru/bitstream/10995/92827/1/978-5-7996-3078-2_2020.pdf
- Hubbard, W. H., Pitkänen, K., Schlumbohm, J., Sogner, S., Thorvaldsen, G., & van Poppel, F. (2002). *Historical studies in mortality decline*. Oslo: Novus.
- Isupov, V. A. (2016). Epidemiologicheskij perekhod v Rossii: vzglyad istorika [The epidemiological transition in Russia: A historian's view]. *Demograficheskoe obozrenie*, 3(4), 82–92. doi: [10.17323/demreview.v3i4.3207](https://doi.org/10.17323/demreview.v3i4.3207)
- Johnson, N. P. A. S., & Mueller, J. (2002). Updating the accounts: Global mortality of the 1918–1920 «Spanish» influenza pandemic. *Bulletin of the History of Medicine*, 76(1), 105–115. doi: [10.1353/bhm.2002.0022](https://doi.org/10.1353/bhm.2002.0022)
- Kashchenko, S. G., & Markova, M. A. (2012). Demograficheskie protsessy v uezdakh Sankt-Peterburgskoi gubernii vo vtoroi polovine XVIII – pervoi polovine XIX vv. Opyt analiza massovoi pervichnoi dokumentatsii ucheta naseleniia [Demographic processes in the Saint Petersburg province in the second half of the 18th – first half of the 19th centuries. Experience in analyzing mass primary documentation]. *Informatsionnyi Biulleten' Assotsiatsii «Istorija i Komp'iuter»*, 38, 55–57.

- Kluesener, S., Devos, I., Ekamper, P., Gregory, I., Gruber, S., Martí-Henneberg, J., ... Solli, A. (2014). Spatial inequalities in infant survival at an early stage of the longevity revolution: A pan-European view across 5000+ regions and localities in 1910. *Demographic Research*, 30, 1849–1864. doi: [10.4054/DemRes.2014.30.68](https://doi.org/10.4054/DemRes.2014.30.68)
- Korkodinova, A., Glavatskaya, E., & Borovik, I. (2016). Brachnye strategii liuteran Ekaterinburga po materialam metriceskikh knig tserkvi sv. Petra (1892–1919 gg.) [Marriage strategies of the Lutherans of Yekaterinburg according to the parish records of St. Peter Church (1892–1919)]. In P. I. Mangilev (Ed.), *Tserkov', bogoslovie, istoriia: materialy IV Mezhdunarodnoi nauchno-bogoslovskoi konferentsii: Ekaterinburg* (pp. 166–172). Ekaterinburg: Theological Seminary.
- Kumo, K. (2017). Changes in mortality: Meta-analysis. In T. Karabchuk, K. Kumo & E. Selezneva (Eds.), *Demography of Russia: From the past to the present. Studies in Economic Transition* (pp. 219–259). London: Palgrave Macmillan. doi: [10.1057/978-1-137-51850-7_7](https://doi.org/10.1057/978-1-137-51850-7_7)
- Markova, M. A. (2016). Smernost' pravoslavnogo naseleniia g. Vyborga po dannym metriceskikh knig XIX–XX v. [The mortality rate of the orthodox population of Vyborg according to the metric books of the 19th–20th centuries]. In V. F. Blokhina (Ed.), *Rossiiia v epokhu politicheskikh i kul'turnykh transformatsii* (pp. 173–175). Bryansk: Kursiv.
- Mazur, L., & Gorbachev, O. (2016). Primary sources on the history of the Soviet family in the twentieth century: An analytical review. *The History of the Family*, 21(1), 101–120. doi: [10.1080/1081602X.2015.1031808](https://doi.org/10.1080/1081602X.2015.1031808)
- Mironov, B. N. (2007). Novaja istoricheskaja demografija imperskoj Rossii (ch. 2): Analiticheskij obzor sovremennoj literatury [New historical demography of Imperial Russia (pt. 2): Analytic review of contemporary literature]. *Vestnik Sankt-Peterburgskogo universiteta*, 2(4), 100–125.
- Palkin, A., & Borovik, I. (2019). Brachnye strategii v edinovercheskoj obshhine Ekaterinburga v nachale XX v. [Marriage strategies of the Yekaterinburg Edinoverie community in the early 20th century]. *Quaestio Rossica*, 7(4), 1311–1323. doi: [10.15826/qr.2019.4.440](https://doi.org/10.15826/qr.2019.4.440)
- Palli, H. (1983). Parish registers and revisions: Research strategies in Estonian historical demography and agrarian history. *Social Science History*, 7(3), 289–310. doi: [10.1017/S0145553200019672](https://doi.org/10.1017/S0145553200019672)
- Patterson, K. D., & Pyle, G. F. (1991). The geography and mortality of the 1918 influenza pandemic. *Bulletin of the History of Medicine*, 65(1), 4–21.
- Pentikäinen, J. & Simoncsics, P. (Eds.). (2005). *Shamanhood: An endangered language*. Oslo: Novus.
- Ransel, D. L. (1991). Infant-care cultures in the Russian empire. In B. E. Clements, B. A. Engel, & C. Worobec (Eds.), *Russia's women. Accommodation, resistance, transformation* (pp. 113–134). Berkeley, CA: University of California Press.
- Sergel, S. (1927). *God Kochevki s lopariami. Ocherki prirody i liudei. [A year of traveling with the Sami. Essays on the people and nature]*. Moscow, Leningrad: Gosudarstvennoe izdatel'stvo.
- Shanks, G. D., & Brundage, J. F. (2012). Pathogenic responses among young adults during the 1918 influenza pandemic. *Emerging infectious diseases*, 18(2), 201–207. doi: [10.3201/eid1802.102042](https://doi.org/10.3201/eid1802.102042)
- Shestova, T. Y. (2017). Razvitie zdravookhraneniya v Permskoy i Vyatskoy guberniyakh v kontse XIX – nachale XX vekov [The development of the public health service in Perm' and Vyatka provinces in the late 19th – early 20th century]. *Historia Provinciae. Zhurnal regional'noy istorii*, 1(1), 24–39. doi: [10.23859/2587-8344-2017-1-1-2](https://doi.org/10.23859/2587-8344-2017-1-1-2)
- Sommerseth, H. L., & Walhout, E. C. (2019). Deaths in a city: A view from the 19th century church registers in Norway. In E. Glavatskaya, G. Thorvaldsen, G. Fertig, & M. Szoltysek (Eds.), *Nominative data in demographic research in the East and the West* (pp. 185–201). Ekaterinburg: Ural University Press. doi: [10.15826/B978-5-7996-2656-3.11](https://doi.org/10.15826/B978-5-7996-2656-3.11)
- Strekalov, D. V., & Strekalova, N. V. (2018). Struktura i tipologija provincial'noj gorodskoj sem'i v konce XVIII – pervoj polovine XIX veka (na materialah Tambova) [The structure and typology of a provincial urban family at the end of the 18th – first half of the 19th century (based on materials from Tambov)]. *Vestnik Tambovskogo universiteta. Serija Gumanitarnye nauki*, 23, 119–130. doi: [10.20310/1810-0201-2018-23-172-119-130](https://doi.org/10.20310/1810-0201-2018-23-172-119-130)
- Strekalov, D. V., & Strekalova, N. V. (2019). Vozrast vstupleniya v brak v provincial'nom gubernskom gorode v konce XVIII – pervoj polovine XIX v. (na materialah Tambova). [The age of marriage in a provincial town at the end of the 18th – first half of the 19th century (based on materials from Tambov)]. *Rus', Rossija. Srednevekov'e i Novoe Vremja*, 6, 276–280.
- Szoltysek, M. (2015). *Rethinking East-Central Europe: Family systems and co-residence in the Polish-Lithuanian Commonwealth. Contexts and analyses*. Bern: Peter Lang.
- Thorvaldsen, G. (2018). *Censuses and census takers. A global history*. London: Routledge.
- Thorvaldsen, G., & Glavatskaya, E. (2017). The three main Western revolutions and their censuses. *Quaestio Rossica*, 5(4), 922–1008. doi: [10.15826/qr.2017.4.263](https://doi.org/10.15826/qr.2017.4.263)

- Troinitskii, N. A. (1905). *Obshchie svedeniia po imperii rezul'tatov razrabotki dannykh pervoi vseobshchei perepisi naseleniia, proizvedennoi 28 ianvaria 1897 goda* [General information on the results of the first All-Russian census of the population, produced on 28 January 1897] (Vol. 1). Sankt-Petersburg.
- Troitskaya I. (1995). *Revizii naseleniya Rossii kak istochnik demograficheskoi informatsii (metodologicheskie problemy)* [The Revisions of the population of Russia as a source of demographic information (methodological problems)] (Ph. D. thesis). Moscow State University, Moscow. Retrieved from https://rusneb.ru/catalog/000199_000009_000096474/
- Ul'yanova, G., & Troitskaya, I. (2016a). Revizskie skazki kak istochnik izucheniia istoricheskoi demografii v istoriografii 1950–1960-kh godov ['Revizskie skazki' as a source of the study of historical demography in the historiography of the 1950s–1960s]. *Vestnik Pravoslavnogo Sviato-Tikhonovskogo Gumanitarnogo Universiteta. II: Istorii. Istoriiia Russkoi Pravoslavnoi Tserkvi*, 68(1), 89–101.
- Ul'yanova, G., & Troitskaya, I. (2016b). Revizskie skazki kak istochnik v istoriografii 1970-kh–2010-kh godov ['Revizskie skazki' as a source in the historiography of the 1970–2010s]. *Vestnik Pravoslavnogo Sviato-Tikhonovskogo Gumanitarnogo Universiteta. II: Istorii. Istoriiia Russkoi Pravoslavnoi Tserkvi*, 71(4), 118–135.
- Vladimirov, V. N., & Sarafanov, D. E. (2013). *Informatsionnye tekhnologii v izuchenii metricheskikh knig (naselenie Barnaula v kontse XVIII – nachale XX v.)* [Information technologies in the study of parish registers (the population of Barnaul in the late 18th – early 20th centuries)]. Barnaul: Altai State University Press. Retrieved from <http://elibrary.asu.ru/handle/asu/403>
- Zabolotnykh, E. A. (2018). Conversion of Jews to the Russian Orthodoxy at the beginning of the 20th century: Microhistorical analysis (based on the materials of Ekaterinburg metric books). In I. V. Krasavin (Ed.), *Mnogomernost' obshchestva: Chelovek v social'nom vzaimodejstvii: 2-j molodezhnyj konvent: Materialy mezhdunarodnoj studencheskoj konferencii 29–31 marta 2018 goda*. Ekaterinburg: Ural University Press. Retrieved from <http://elar.urfu.ru/handle/10995/61741>
- Zabolotnykh, E. A. (2020). Metricheskie knigi Bogojavlenskogo prihoda Ekaterinburga: Kritika i informacionnye vozmozhnosti istochnika [Ekaterinburg Epiphany Church's parish books: Source criticism]. In *Materialy Ural'skogo istoriko-arhivnogo foruma, posvjashhennogo 50-letiju istoriko-arhivnoj special'nosti v Ural'skom universitete* (pp. 49–54). Ekaterinburg: Ural University Press. Retrieved from <https://elar.urfu.ru/handle/10995/92880>

LINKS

A System for Historical Family Reconstruction in the Netherlands

Kees Mandemakers	International Institute of Social History, Amsterdam & Erasmus University Rotterdam
Gerrit Bloothoof	Utrecht University & Meertens Institute, Amsterdam
Fons Laan	International Institute of Social History, Amsterdam
Joe Raad	LISN, CNRS (UMR 9015), University of Paris-Saclay
Rick J. Mourits	International Institute of Social History, Amsterdam
Richard L. Zijdeman	International Institute of Social History, Amsterdam & University of Stirling

ABSTRACT

LINKS stands for 'LINKing System for historical family reconstruction' and is a software system to link nominal data from the Dutch archives and ultimately reconstruct historical individuals and families. We present the background and philosophy of this matching system and explain its data structure and functioning. Currently the core data of the LINKS system consists of indexed civil certificates. These certificates are available from 1812 — the start of the Dutch Vital Registration — until the year they are confidential based on privacy laws. For more than 20 years, thousands of volunteers have been working to build this index, which contains not only the names of newborn, married and deceased persons, but also the names of their parents, places of birth, ages and sometimes their occupational titles. The software system LINKS includes the standardization of all input before linking, nominal record linkage procedures and identification of all unique persons involved in the system. All processes are repeatable and a strict distinction is maintained between source data, standardized, linked and enriched data and released data. Moreover, LINKS also informs archives about all kinds of errors and inconsistencies found during the cleaning and matching process. We will discuss two matching systems, the first is the original querying system that runs within a MySQL database environment and the second is a newly developed system, called burgerLinker, which is based on knowledge graphs and which is designed as a system that can be used independently from LINKS and is made available as open source software. Finally, we present the most important releases of LINKS data so far: two national releases that link birth and parental marriage certificates, creating families and pedigrees and an integrated dataset of persons, families and family trees in four provinces.

Keywords: Nominal record linkage, Historical population data, Civil certificates, Historical demography, Family reconstitution, Genealogical data

DOI article: <https://doi.org/10.51964/hlcs14685>

© 2023, Mandemakers, Bloothoof, Laan, Raad, Mourits, Zijdeman

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

LINKS stands for 'LINKing System for historical family reconstruction' and is a software system to link nominal data from the Dutch archives and ultimately reconstruct historical individuals and families. Such a reconstruction is indispensable for all scientific work dealing with people in the past, and also facilitates the work of genealogists enormously. Given privacy constraints, the reconstruction of the Dutch family network population is possible until about 100 years ago, while it could go back in time as far as the 17th century, depending on locally available sources. In this paper, we present the background and philosophy of this rule-based matching system and explain its data structure and functioning.

Currently the core data of the LINKS system consists of indexed civil certificates. These certificates are available from 1812 — the start of the Dutch Vital Registration — until the year they are confidential based on privacy laws. This limitation depends on the type of the certificate and is respectively 100, 75, and 50 years for birth, marriage, and death certificates.¹ For more than 20 years, thousands of volunteers have been working to build this index, which contains not only the names of newborn, married and deceased persons, but also the names of their parents, places of birth, ages and sometimes their occupational titles. In 2022 the index contained over 125 million person names and is continuously growing not only because new civil certificates become public each year but also because existing gaps are filled. All digitized data are publicly accessible through WieWasWie ('WhoWasWho', see <https://www.wiewaswie.nl>), based at the CBG Center for Family History (<https://www.cbg.nl>).

When designing LINKS three requirements were formulated to ensure both a successful reconstruction of historical persons and dissemination of releases with linked data: a) standardization of all input before linking, b) development of nominal record linkage procedures and c) identification of all unique persons involved in the system. All processes are repeatable and the database maintains a strict distinction between A) source data, B) standardized, linked and enriched data and C) released data (Mandemakers & Dillon, 2004). Moreover, LINKS also informs archives about all kinds of errors and inconsistencies found during the cleaning and matching process.

LINKS is based at the International Institute of Social History (IISG) as part of the HSN databases (HSNDB). Beginning in 2006, releases were disseminated from the indices of the civil certificates, mainly matched marriage records. These releases were initially known as the GENLIAS datasets, before the name LINKS was adopted. LINKS started in 2010 as a spin-off of the Historical Sample of the Netherlands (HSN) (Mandemakers & Kok, 2020), financed by the NWO CATCH program.² The project was a cooperation between the IISG, Utrecht University, the Meertens Institute and the Leiden Institute of Advanced Computing. For more information about the LINKS project, see <https://iisg.amsterdam/en/hsn/projects/links>. Nowadays, LINKS is part of HSNDB, the IISG system of databases for historical and contemporary research (see <https://iisg.amsterdam/en/hsndb>).

In the next three sections of this paper we explain the workflow and processes of the LINKS system; in the last two sections we describe the construction of the major releases. Section 2 concentrates on the sources that form the basis of LINKS. Section 3 focuses on the cleaning of the imported data from WieWasWie and on the feedback given to the archives that provide the WieWasWie data. In Section 4 we discuss two matching systems, the first is the original querying system that runs within a MySQL database environment and the second is a newly developed system, called burgerLinker, which is based on knowledge graphs. BurgerLinker is designed as a system that can be used independently from the LINKS system and made available as open source software, so that it can be used freely for all kinds of nominal data.³ In Section 5 we evaluate the outcomes of different matching strategies applied on the Zeeland marriage certificates. In Section 6 we present the most important releases of LINKS data so far: two national releases that link birth and parental marriage certificates, creating families and

1 [Burgerlijk Wetboek](https://www.wetten.overheid.nl/BWBR0002656/) ('Dutch civil code'), article 1:17A. Retrieved 5 January 2023 from <https://wetten.overheid.nl/BWBR0002656/>

2 CATCH stands for Continuous Access to Cultural Heritage and is a program of the Dutch Research Council (NWO). In this program researchers and heritage managers worked together to make heritage data more accessible and develop instruments to enable heritage managers to work more efficiently. LINKS was one of the 12 projects that were granted. The programme started in 2004 and ran till 2014, for more information see <https://www.nwo.nl/en/researchprogrammes/continuous-access-cultural-heritage-catch>

3 See <https://www.github.com/clariah/burgerlinker>

pedigrees and an integrated dataset of persons, families and family trees in four provinces. The paper ends with a summary and conclusion.

2 DATA FROM THE DUTCH CIVIL REGISTERS AND THE LINKS WORKFLOW

First attempts with record linkage in historical demography were done with data from church records and civil registers. Louis Henry is the well-known founder of a methodologically grounded way of linking this kind of records. Together with Michel Fleury he developed a form to create and record family reconstitutions (Henry & Fleury, 1956; Séguy, 2016). The first datasets of this kind were limited to the parish area. Examples are the reconstitution of 34,812 families in 39 French parishes from the period 1640–1829 (Séguy, 2001) and the database constructed by the Cambridge Group for the History of Population and Social Structure for 26 parishes in England and Wales over the period 1580–1837 (Wrigley, Davies, Oeppen, & Schofield, 1997). With the growth of computing power and expertise within the field, historical reconstructions are now available for a myriad of countries and the scope is only increasing (for an overview, see Mandemakers, 2023; Song & Campbell, 2017). After the parish level whole nations came into view. In Québec, projects started with the aim to reconstruct the whole population from 1621 onwards (Dillon et al., 2018; Nault & Desjardins, 1989; Vézina & Bournival, 2020). In France, Dupâquier and Kessler (1992) collected a sample of 40,000 marriage certificates from all over France based on the letter combination TRA and linked them into pedigrees. Subsequently other researchers added other data to this basic construction such as birth and death certificates, military registers and data from hereditary tax and military registers (Bourdieu, Kesztenbaum, Postel-Vinay, & Tovey, 2014). Based on the Dutch civil records, LINKS is a continuation along the path set out especially by these French historical demographers.

Civil registration was introduced in the Netherlands in 1810, as a consequence of the annexation by the French Empire. The Code Napoléon provided for the compulsory, standardized recording of vital events in certificates. The certificates had to be drawn up in the municipality where the vital event occurred. Most Dutch municipalities introduced civil registration over the course of 1811. However, since the Dutch province of Limburg and the south of Zeeland (Zeeuws-Vlaanderen) were annexed by France in 1796, civil registration for these provinces was introduced in that year (Vulmsa, 1988).

All certificates of birth, marriage, or death ever made in the Netherlands are still available, as each certificate was made in duplicate and stored in books for safekeeping. At the end of each year, one civil registry book remained in the municipality and the other was sent to the provincial courts. The registrars had to note the name, age, occupation and municipality of residence of the informants and witnesses. This information assured the correct identification of these individuals. In Dutch birth certificates we find the names, address, ages and occupations of the parents in addition to data on the newborn. Death certificates provide last residence, age and final occupation of the deceased and data on the spouse(s) and parents, including occupational titles if they were still alive. The information concerning the parents was officially less detailed, as age was not required, but nevertheless its registration was widespread. The marriage certificates give information on the occupations, illiteracy (absence of signature) and places of residence of the bride, the groom, their parents and the (usually four) witnesses, who were relatives or friends of the marrying couple about half the time (Mandemakers, 2000; Vulmsa, 1988; for an exhaustive list of all information found in the certificates, see Mourits, van Dijk and Mandemakers (2020), p. 44, table 1). Data from civil certificates are non-dynamic, which implies that they are only valid at the date of the events of birth, marriage and death. This is fundamentally different from sources that offer a more continuous stream of data like the population register which is used by the Historical Sample of the Netherlands (HSN). For a systematic comparison of the value and use of both sources, see van den Berg, van Dijk, Mourits, Slagboom, Janssens and Mandemakers (2021).

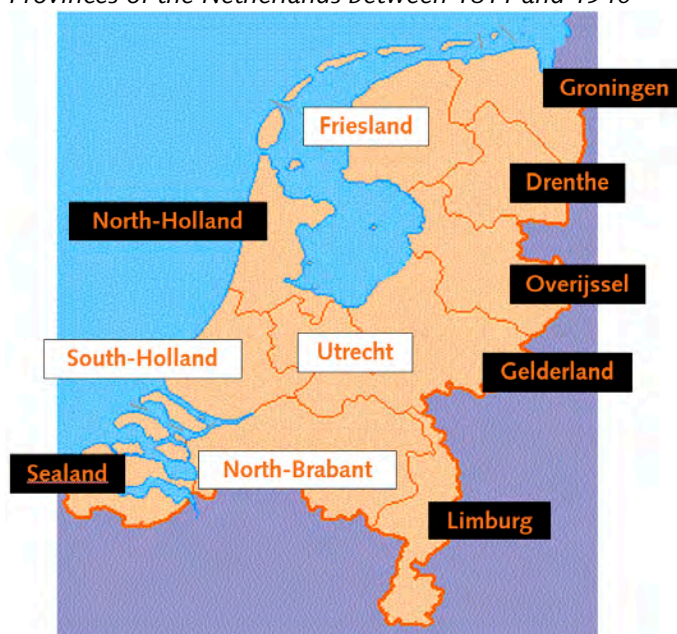
Since the early nineties of the previous century, hundreds of volunteers have been working on the indexation of all names and dates from these certificates. This indexing was embedded in the already existing practice in which local and regional archives organized volunteers to help disclosing archival collections. Identifying persons and buildings from old photographs and films was a favorite exercise. It was probably the city archive of Amsterdam that started the first volunteering project with population data by starting the data entry of the population register 1850–1853. Soon after, an initiative by the National Archive called GENLIAS started to index all marriage certificates in a consistent way. Around

1995, provincial archives were firmly encouraged to cooperate and recruit the volunteers required for indexing. In the beginning indexing was done using the original sources, but from the year 2005 onwards data entry from scans became the norm and later data entry also became web based. Volunteers could work at home, which extended enormously the base of volunteers. The organization also professionalized in the sense that data entry was done more and more by private companies offering data entry programs embedded within platforms. This also created a kind of community to advise the volunteers on problematic issues and to monitor the progress of a specific job. Important companies are *Vele Handen* ('Many Hands') and *Het Volk* ('The Crowd') that presently organize about 35 different projects, ranging from the indexing of notarial deeds, population registers to indexing photographs.⁴

A new platform called *WieWasWie* was constituted to present the information from the Dutch local and regional archives and deal with the technical challenges in the presentation and search possibilities of the indexed data and linked scans. *WieWasWie* is maintained by the Dutch Family Center and offers a central point for all archives to present their indexes and to make searches for persons on a national scale possible (<https://www.wiewaswie.nl/en/>). Besides names and dates, indexed information differs between archives, as *WieWasWie* is designed as a decentralized system. The participating archives can make their own decisions as to which data are entered into the index, but always included are the type, date and municipality of the event as well as the first names and family names of the persons involved (child/parents, bride/groom/parents or deceased/parents/partner) and usually age at the event for the deceased, bride and groom. Witnesses are seldom included. Occupational titles were systematically entered for the marriage certificates in seven out of eleven provinces, see Figure 1, adding up to about 60% of all certificates. This percentage is much lower in the case of death and especially birth certificates.

For privacy reasons, certificates are made public with a delay of 100 years (birth certificates), 75 years (marriage certificates) or 50 years (death certificates), so in 2018 certificates were available until 1918, 1943 and 1968, respectively. In practice, the delay is up to 5 years longer as most archives do not update their indexes annually. Table 1 presents the level of indexation in September 2018. It shows that the marriage certificates are almost completed. Lagging behind are the birth and to a smaller degree, the death certificates. Currently about 85% of the indexing has been done. In all, 27 million civil certificates were digitized, containing information on about 120 million person mentions. However, archives are quickly catching up, and we expect countrywide coverage in about 5 years.

Figure 1 Provinces of the Netherlands between 1811 and 1940



Explanation: Provinces for which occupational titles are available have a black frame; in the province of South-Holland the cities The Hague and Leiden are a positive exception since they also included the occupational titles in the index. In the case of brides and parents one often finds the term "without occupation". Occupations of deceased parents are not mentioned at all.

4 Another article in this special issue concerns the slavery registers of Suriname of which the data entry was also done by volunteers (van Galen, 2019; van Galen et al., forthcoming).

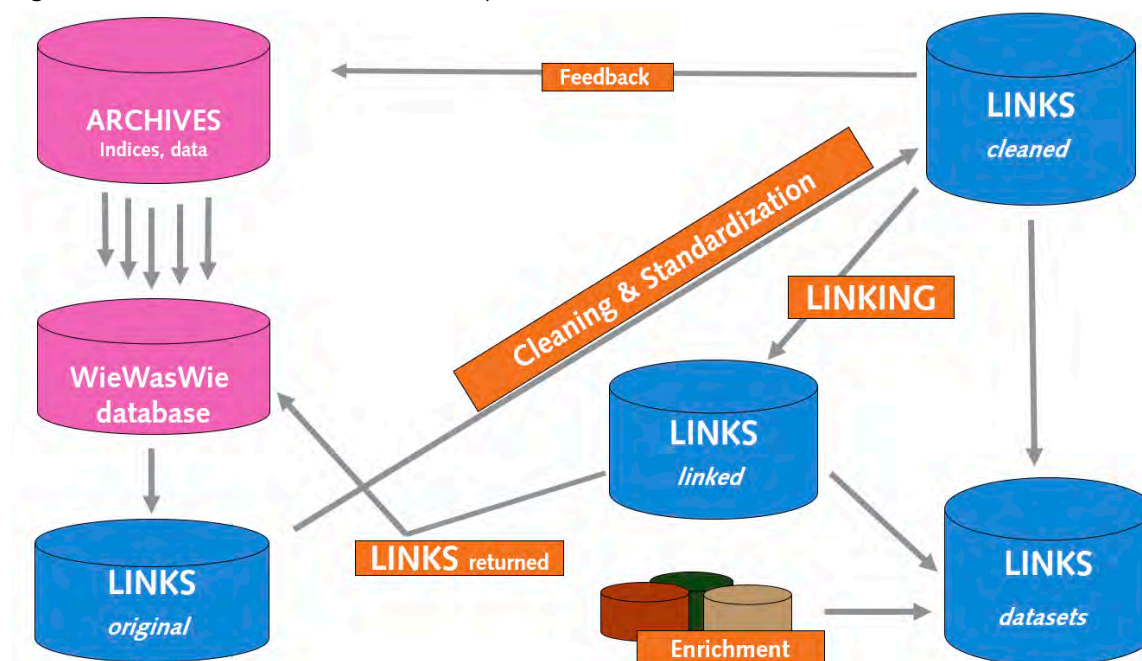
Table 1 *Number of the publicly available and indexed civil certificates (in millions), the Netherlands, September 2018, WieWasWie*

	Indexed	Public	% Indexed
Birth certificates	10.4	14.2	73.2
Death certificates	12.1	13.2	91.7
Marriage certificates	4.5	4.8	93.8
Total	27.0	32.2	83.9

Sources: The numbers of total events before 1850 were provided by van der Bie and Smits (2000), except the marriages 1812–1839 which were estimated; the number for the period from 1850 onwards were provided by the Historical Database of Dutch Municipalities (HDNG; Mourits, Boonstra, Knippenberg, Hofstee, & Zijdemann, 2016). The number of indexed certificates was calculated from the data that were gathered by LINKS in September 2018.

The LINKS system is designed in a generic way and can handle nominal data from all kinds of sources. But in the development phase, we used data from the civil registration available in WieWasWie to create datasets with linked certificates in an enriched way and made them available to the scientific community. We also created reconstructions of life courses and families for separate regions in the Netherlands. In Figure 2 the general outline of the LINKS workflow process is sketched. Data are delivered by the regional and city archives to the WieWasWie-system. The ownership and responsibility for the content rest with these archives. LINKS harvests the data from WieWasWie in the LINKS *original* database as soon as new releases of data are added to WieWasWie. From LINKS *original* the data are cleaned and standardized and subsequently added to the LINKS *cleaned* database. We give feedback to the WWW community on the errors and problems we found. These cleaned data form the basis of the different matching procedures. The resulting links are stored in LINKS *linked*, while the links between the certificates are also returned to the Dutch Family Center to be published on the website of WieWasWie. In a further stage the data from LINKS *cleaned* and LINKS *linked* are combined into LINKS *datasets* with, for instance, pedigrees and families, and enriched with geographical data and occupational coding. These data sets are released for scientific research.

Figure 2 *Workflow of the LINKS system*



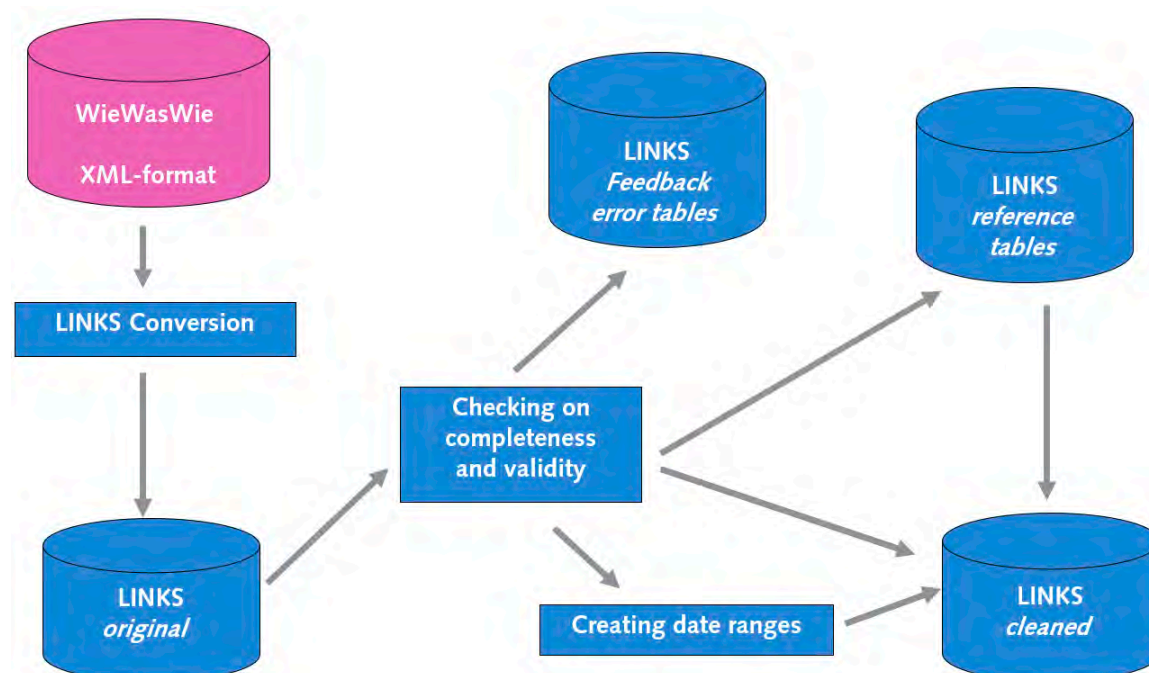
3 PREPARING THE DATA FOR MATCHING

3.1 CONVERTING, CLEANING AND STANDARDIZATION

The data from the WieWasWie database is made available in XML format. The LINKS system consists of several MySQL databases based on two tables: one table for information on the registration of events, and one table with information on all persons involved in the events.

Figure 3 focuses on the part of the workflow that processes the data from the WieWasWie dataset to the LINKS *cleaned* database. The first step is to import the data from XML into MySQL tables, and to distinguish the different types of events and define the corresponding roles of the individuals involved. This is a source-dependent operation put in place for every document type, currently birth, marriage and death certificates, but extensions of these scripts are easy to provide. In this conversion the character type is also converted to UTF-8 and all diacritics are changed into basic characters (e.g., é, ë, and è become e). The result is stored in the database LINKS *original*.

Figure 3 LINKS workflow from original to cleaned data



Once the data is stored in the LINKS *original* database, a cleaning process is started which results in the LINKS *cleaned* database. The cleaning process checks all data on completeness and validity. Completeness concerns the presence of compulsory information, such as including all roles belonging to a type of registration. For example, a marriage certificate should include at least a bride and a groom and not include roles exclusively belonging to a birth or death certificate. Other checks test on duplicated roles in a certificate, or whether each person has a first name and a family name which is essential for nominal linking. Validity concerns the consistency and range of dates; all dates should be in a range that is in agreement with the properties of the source type. For example, the date of a birth certificate cannot be earlier than the date of the birth itself, or a groom or bride should be at least 14 years younger than their parents. Incomplete, inconsistent or invalid data are reported in a systematic way (see Appendix A for an overview of error messages).

Data cleaning also involves standardization, for example avoiding variation that is not essential to the meaning of the information. This may concern spelling variation in place names, occupational titles, variation in the writing of dates, the writing of abbreviations in full, and so on. In some cases the spelling of first names or family names is also standardized. Standardization procedures in LINKS apply so-called reference tables. The most important ones are those for family names, first names, ages, locations, occupational titles, sex, and civil status. For every type of information, the system verifies whether the content is already present in the relevant reference table. When the original value is already included, a code is assigned describing the validity of the entry under three possible values:

One for "valid and standardized", one for "not valid, but clear enough to be standardized" and "not valid and not standardized." Corresponding standard values are written to LINKS *cleaned*. If the content is not known in the reference table, it will be written in LINKS *cleaned* and a new record will be made in the reference table where it awaits standardization. After a new round of standardization, cleaning is run again on every field. If the value is considered invalid, it is included in the error table and reported to the originating archive (see Appendix A for these messages).

Reference tables are part of the existing standardization schemes in the HSNDB domain. With each new release, all new values in the LINKS database are standardized and coded in a manual and/or semi-automatic way and added to the reference table after expert review. Table 2 gives an overview of the content of the main reference tables as of July 2021. The first column gives the number of original values, the second column the number of standardized values, the third column the number waiting to be standardized, and the fourth column the values that were considered invalid. The fifth column presents the resulting unique values. These tables are kept in one system together with other databases that form the HSNDB environment. Additionally, there are smaller reference files for data such as suffixes, aliases, sources, roles and source types. Civil status quite often also implies an indication for the sex of a person (bride, widower etc.), hence a combined table Status_sex was developed. The table Prefixes includes text that may precede a family name as part of the family name (for example "de Boer" instead of only "Boer", which is quite common in the Netherlands), or as a title. Both possibilities are separately standardized.

Table 2 *LINKS reference tables with number of original and standard values, 1–7–2021*

	Originals	Standardized	Not yet standardized	Not valid	Unique standards
Family names	846,860	797,519	49,217	124	508,877
First names	300,299	263,146	36,991	162	238,205
Prepieces (titles and prefixes)	6,264	2,201	1,140	2,923	344
Ages (days/weeks/months/years)	75,318	60,151	0	15,167	2,576
Locations	566,068	150,182	412,148	3,738	8,736
Occupations	319,215	276,665	37,860	4,690	83,610
Status_sex (combined table)	707	674	18	15	23
Religion	2,630	2,629	0	1	117

Based on knowledge of name spelling and after experimental matching with marriage certificates of Zeeland (see Section 5.2), we decided to apply a simple initial standardization for both first and family names by replacing all "ch" by "g", "c" by "k", "ph" by "f", "z" by "s", and "ij" by "y". Furthermore, family names or first names occurring in the whole dataset less than three times were considered to be spelling errors and were standardized to the most frequent corresponding name with a Levenshtein distance of 1. We considered them by definition as spelling mistakes, for example "Gerrjt" with frequency 2 was standardized to "Gerrit". A family name as "Bakkr" with a frequency less than two was standardized as "Bakker". Through these two operations the number of unique standard family names dropped by about 35%, and the number of unique standard first names by about 10% (see Table 2). Since adding unique names will result in more pairs being compared in an exponential way, limiting the number of spelling variants of names is quite important in reducing processing time.

The standardization of other variables is also processed in a semi-automatic way. For ages, the standard is a combination of four values for days, weeks, months and years. Currently, there is a high number of non-accepted values, as some archives entered dates of birth in the age fields rather than the age mentioned on the certificate. Almost 320,000 different original occupational titles are about 80% standardized now, resulting in about 83,000 unique standards. New versions of reference tables are released periodically, see for example Mandemakers et al. (2020) with the latest release of occupational titles (n=281,355) including standard values and corresponding HISCO, HISCAM and HISCLASS classifications (Lambert, Zijdeman, van Leeuwen, Maas, & Prandy, 2013; van Leeuwen & Maas,

2011; van Leeuwen, Maas, & Miles, 2002). Locations are standardized with respect to municipality, province/region and nation and enriched with geo-referential codes (Huijsmans, 2020). The region is a more general category used to cover regional levels above the municipality level, especially outside the Netherlands for example states in the USA, islands in the East-Indies, etc.

The output of the LINKS cleaning system will lead to feedback to the partners of WieWasWie. This may involve general suggestions to improve the quality and uniformity of the data, but also offering specific instructions to verify the data in certain certificates. Original entries which are not accepted as valid data are included in the error archive and reported to the relevant archive. In Appendix A we give an overview of all types of errors reported by LINKS and sent to the archives. Presently there are 114 different types, which resulted in a total of over 3 million messages so far (of which 1.9 million for the death certificates, for about 900,000 because of an inadequate age and 400,000 because of lacking or insufficient firstnames).

3.2 TIME RANGES FOR BIRTH, MARRIAGE AND DEATH

For all persons in each certificate, the time range in which they likely were born, married or died was calculated. These ranges limit the number of false positives and are also used to decrease processing time. The estimation of the minimum and maximum years of the range is based on six features:

- 1 The type of a certificate;
- 2 The role of a person in a certificate;
- 3 The age of a person at the event (if known);
- 4 Whether a person is alive or not at the event (if known);
- 5 The age of a related person in the certificate;
- 6 Preset ranges for certain life events.

The last feature is operationalized by the following rules:

- 1 A woman will give birth to children at an age between 14 and 50 years;
- 2 A man will father children at an age between 14 and 100 years;
- 3 Children are born in a legitimate state, i.e. parents are married at child birth;
- 4 Persons will not become older than 110 years of age;
- 5 Difference in age between partners will be maximally 66 years;
- 6 Maximal age of marriage for a woman is 90 years;
- 7 Maximal age of marriage for a man is 100 years.

For example, consider the mother of the groom in a marriage certificate from 1888. We know from the certificate that the groom is 25 years old. That implies that given a fertile period of the mother of 14–49 years she cannot be older than $25+50=75$ years and not younger than $25+14=39$ years, so the mother should be born between 1813 and 1849. We can also calculate the range of her year of marriage which is between $1813+14=1827$ and $1849+14=1863$, which should also be the range of the year of marriage of her husband, the father of the groom. If the mother was present at the marriage of her son, we know that she died between 1888 and $1849+110=1959$ and if she was mentioned as deceased, we can expect that she died between $1888-25=1863$ and 1888.

All calculation rules are defined in a specific table in which we may change our assumptions of minimum and maximum ranges. For example, we may change the range of the maximal age of a married woman as soon as we find someone who married after the age of 90. More critical is the assumption of giving birth in the age range from 14–50 years; since births below the age of 16 or above 47 are rare. We could change this into a range of 16–47 years, balancing between missing a few matches or generating false ones. Even more critical is the condition of "born in a legitimate state". A previous study on a sample of the population of the province of Noord-Holland showed that on average during the 19th century this is not the case for about 5% of all first-born children (Kok, 1991, pp. 46–48). From the birth certificates of children of all birth orders we found that in 1.5% of the cases no father was mentioned (Mandemakers & Laan, 2020a). Because these children were not recognized by their father upon birth, their biological relation is debatable. However, we might still be dealing with correct matches, when the father later appears as a legal father on the child's marriage certificate.

To explore how many links we could have missed by assuming "born in a legitimate state", we experimented with larger margins in the linking of marriage certificates, putting the potential wedding range twenty years earlier. We found for the whole of the Netherlands about 281,000 extra linked marriage certificates of which about 38,000 with a margin of one year, 62,000 with two to three years and 52,000 four to five years. Since for larger margins the percentage of exact matches dropped from 42% to 12% or lower, we decided that five years should be considered as a limit for accepting matches. In a second test we took a small sample of 24 *exact* matches from this range of maximum five years to check if these children were formally acknowledged by the father on the birth certificate and/or on the parental marriage certificate. This was true in 100% of the cases.⁵ Since it is easy to implement changes in the ranges, we intend with new releases to relax this requirement into "born in or five years before a legitimate state". Although this may result in relatively more false links, the researcher can make his own decisions on the basis of information about the time lag and the quality of the matches (especially the father link since his name is not always originally included in the birth certificate).

A birth certificate includes three roles (child, mother and father), three different conditions for calculating ranges (age of the involved person known, no age known, or known to be alive or not), for three events (birth, marriage or death). This results in $3 \times 3 \times 3 = 27$ cases to take into account when calculating minimum and maximum time ranges for all three roles in a birth certificate. For the marriage and death certificates there are respectively 54 and 36 different conditions. The calculation of age ranges is even more complicated since in some cases interdependencies exist between the different procedures which have been solved by developing several specific functions that overrule the outcomes of the initial calculations. In the previous example, this concerns the theoretical marriage range of the father (1792–1863) which is limited by the marriage range of the mother (1827–1863) or the minimum range of death which is always defined by an event in which a person is registered as being alive.

4 DESIGN OF THE MATCHING SYSTEMS

4.1 MATCHING APPROACHES

The linkage problem posed by the LINKS database is the identification of individuals and their family relations on the basis of multiple mentions in historical civil records. This process is complicated because names are seldomly unique identifiers for persons, even though in the Dutch civil administration everyone (also women) keeps the first name and family name given at birth during life time. Still, the same person may occur with different (spellings of) names, and a single name may refer to multiple persons. Therefore, for the identification of an individual (ego), related actors are needed, notably the parents and partner(s). The combination of multiple names and time ranges for birth, marriage, and death has a high probability of leading to a unique identification. Functional relational combinations are ego and partner, ego and mother (mentioned at birth, marriage and death of ego), ego, father and mother (mentioned at birth, marriage and death of ego), and ego, parents and partner (mentioned at marriage and death of ego).

The combination of the ego and partner forms the backbone of our family reconstruction as they are mentioned as bride or groom in their own marriage certificate, and can be linked to the marriage of their children where they are mentioned as parents. Links to their mentions as parents in the birth certificates of their children makes it possible to form families, while links to their mentions in death certificates complete their life history of vital events. In several matching operations combinations of three or four persons could be used to match certificates. In principle, by using more than two persons these matches show less ambiguous results than matching on only two persons (ego and mother). Once these relationships have been established the full family reconstruction is realized in a post-matching stage (see Section 6.3).

In the first instance a matching program was developed using SQL queries in a MySQL database. In Section 4.2 we will explain this SQL-based system. However, this system was relatively slow.

⁵ We sampled 12 cases for the father line and 12 for the mother line, each divided in three groups of four, one for 0–1 years before the marriage, one for 2–3 years and one for 4–5 years. We also checked eight parental marriage certificates of children who were born 6–10 years before their mother married. These children had in less than 50% of all cases, the same person as father as the one appearing on the marriage certificate of their mother (and appearing as (a false) father on their own certificate).

Therefore, in order to deal with the ever-increasing scale on which civil certificates were matched, a much faster system was developed by Joe Raad. This system called "burgerLinker" (*burger* meaning 'citizen') mostly applies the same matching rules, but matches compressed knowledge graphs rather than MySQL data. In Section 4.3 we will explain this graph-based system.

4.2 LINKING WITH SQL-QUERIES

4.2.1 PREMATCH TABLES

For the matching of family names and first names we use Levenshtein edit distances. We also tested the Jaro-Winkler algorithm which gives a stronger weight to the first characters of a name (Schraagen, 2014). However, this algorithm did not provide better results. This is probably the effect of the very good quality of the names in the certificates, indicating that the last half of a string is not much more vulnerable to spelling mistakes than the first half.

To speed up the matching process, we make use of prematch tables, one for first names and one for family names. These tables include all combinations of family names or first names within a certain maximum Levenshtein distance. Matching two names is thus simplified since it is possible to find relevant combinations in look-up tables. Because the Levenshtein distance is influenced by the length of both names, the maximum distance was made dependent on the length of the shortest name of a pair. This relation is given in Table 3, which presents the number of matched pairs of names for both first names and family names. We created two prematch tables with a distinction on the way Levenshtein distances and lengths are combined: one with relatively free requirements and a stricter one. In order to save processing time in the case of the freer one, we blocked on the first character. As one can see in Table 3, this has the effect that for a minimal length four or shorter, the freer variant has less matched pairs than the strict one. However, for name lengths of five and higher, the freer one results for the first names in a total of 15.22 million pairs to be looked up, whereas the strict one ends with 7.12 million. We see a similar mechanism with the family names.

Working with prematch tables has the big advantage that during the matching process it is not necessary to calculate Levenshtein again and again for the same pair of names. By excluding pairs with high Levenshtein values from these tables we make the matching process even simpler. So, we limit the pre-match tables to relatively low Levenshtein values that could lead to matchable results. Before the start of the matching process the user needs to put a limit on the accepted Levenshtein variance and to choose which table must be used by the system.

Besides the reference tables with spelling variants of names, we also developed so-called "Root name" tables, both for first names and family names. The idea is that two names could refer to the same individual while they have a large Levenshtein distance. Language is an issue, where "William" "Guillaume" or "Willem" may denote the same person (Bloothoof et al., 2020; Oosten, 2008).⁶ The same occurs with abbreviations or short forms such as "Jan" which originates from "Johannes" with a Levenshtein distance of 6. We make use of existing tables with root names (Bloothoof & Schraagen, 2015). We also intend to develop new variants by checking combinations of non-matching names in situations where all name elements (minimal two first names and two family names) have a match except one element.

However, we are not sure how to use root name matching. First experiments show that when we use root matching above the existing matching with spelling variants, we get about 5% more matches but at least half of them proved to be false. So, this requires additional decision rules to make distinctions between true and false positives.

It is also possible to include the third element in Dutch name structure: the prefix such as "de" in "de Boer". So far, we have ignored the prefix in the linkage process, since it adds little value to the uniqueness of a name.

First names may also contain more than one element, so-called multiple names, e.g., "Cornelia Theresa Antonia Maria." In the look up tables we handle each name separately. This makes it possible to match only the first or the first two names of a multiple name or matching one part of a multiple name with any part of the other name or more variations.

6 See the CLARIAH financed NAMES project, <https://taalmaterialen.ivdnt.org/download/names-corpus/>

Table 3 *Number of pairs of names, with required Levenshtein distance in relation to the length of the shortest name*

Maximum accepted Levenshtein distance	More free application with first character blocked			More strict application		
	Minimal length of the shortest name	Frequency in millions		Minimal length of the shortest name	Frequency in millions	
		First names	Family names		First names	Family names
0	1	0.23	0.53	1	0.23	0.53
1	2–4	0.77	1.36	2–4	1.00	1.83
2	5–7	6.12	12.23	5 or longer	7.12	14.46
3	8	4.82	7.21			
4	9 or longer	4.28	8.36			

Explanation: Frequency numbers are reciprocal.

First names may also contain more than one element, so-called multiple names, e.g., "Cornelia Theresa Antonia Maria." In the look up tables we handle each name separately. This makes it possible to match only the first or the first two names of a multiple name or matching one part of a multiple name with any part of the other name or more variations.

To speed up the matching process, all (standardized) family names and first names were replaced by numbers (one unique number for each name). In another step these tables with names and numbers were enriched with the frequency of each name in the whole dataset. By way of these frequencies, it was possible to direct the matching algorithm in such a way that names with the lowest frequencies were compared first. In this way the selections of potential matches were kept as minimal as possible to save computing time.

4.2.2 MATCH INSTRUCTIONS BY WAY OF A TABLE

The various choices that can be made in the matching procedure are included in the LINKS system by way of a table, called *Match_Process*, which includes the settings of all parameters that govern the matching process. See the scheme in Table 4 summarizing all parameters that can be set for each linking process.

Each record in the *Match_Process* table defines a specific matching procedure. To control the number of comparisons for matching, and by this the processing time, it is also possible to limit the time window for comparisons in a dynamic way.

The *Match_Process* table first defines the two sources to be matched and a time window within which the matching should occur. For example, in marriage to marriage matching the time window could be set in such a way that the parents found in a marriage certificate of a child only match with their own certificate 15 to 75 years before. Secondly, the period of matching can be divided into subperiods to reduce the number of comparisons, and by this processor time, for example to a time range of 10 or 20 years. In case of a time range of twenty years the first 'window' of matching includes the period 1811–1830, in which certificates are matched with parental certificates from the period 1736–1815. In the second window the certificates from 1831–1850 are matched with those from 1756–1835, etc.⁷ By constructing different time windows and matching criteria, relatively small batches are created which are processed in a simultaneous way with 20 to 30 processors.

Another parameter controls the way multiple first names are handled. Multiple first names are not always complete or written in the same order, which especially affects the names of parents. For this reason, there are three options in linking multiple first names, a) on the basis of the first two names, b) only the first name or c) only one of all names (which is a very free method, for example "Johannes Christiaan" will match with "Christiaan Arnoldus Petrus Maria").

⁷ There are no civil certificates before 1811 except in two regions, but the system works with vast ranges that also must cover later periods, for example 1931–1950 compared to 1860–1935.

Table 4 *Parameters to be set in the matching process of two sources*

	Settings source 1	Settings source 2
Type or sources (in all combinations)	Birth, Marriage or Death	Birth, Marriage or Death
Type of archive	All archives or a specific selection	All archives or a specific selection
Definition of ego	Role name of ego	Role name of ego
Combination of roles to be matched (couples, triples or quadruples)	Bride and groom (M) Child and mother (B, M, D)	Mother and father (B) Child and mother (B, M, D)
	Child, mother and father (B, M, D) Child, mother and partner (M, D)	Child, mother and father (B, M, D) Child, mother and partner (M, D)
	Child, mother, father and partner (M, D)	Child, mother, father and partner (M, D)
Use time range	Per combination of type of role and source to be set on/off	Per combination of type of role and source to be set on/off
Window of matching	Defining the start and end year of the matching window in a sequential way	Defining the start and end year of the matching window in a sequential way
Settings of source 1 and source 2 or the same		
Familyname	Type look up table	
Familyname	Maximum Levenshtein level	
First name	Type look up table	
First name	Maximum Levenshtein level	
First name	Coding how different components of multiple first names are to be handled	

Explanation: Possible combinations of roles are dependent on the sources and are indicated with B (Birth certificate), M (Marriage certificate) and D (Death certificate).

Cases matching pairs of two persons involve four name elements to be compared. The order of matching is done in such a way that the first comparison is made for the element with the lowest frequency. Secondly the other name elements are taken into consideration. E.g., in case one family name is "Bakker" with a frequency of 186,231 and the other one is "Zeldenrust" with a frequency of 1,059 the comparison is limited to the selection including "Zeldenrust". The last step is that the outcome is compared with the time ranges as defined for each person (which will always be a subperiod from the set time window for the overall matching, see Section 3.2). So, the date ranges are used as the last step of the process for accepting a match between two certificates or not. The outcome of each comparison is no match, one match or multiple matches. The looser the matching criteria the more matches and multiple matches will be created. To further speed-up the process, a subsequent comparison step is only performed if its predecessor succeeded.

4.3 LINKING WITH BURGERLINKER

4.3.1 BACKGROUND

The matching software within the MySQL environment was used to match records from the Dutch province of Zeeland and other relatively small areas. On the basis of the matching results, family reconstitutions were created (see Section 6). But on a national level, where tens of millions of certificates needed to be matched, the SQL environment proved to be relatively slow, which led to run-time problems and compelled splitting the job into sub-jobs. An alternative to the LINKS software was needed to speed up matching and to create a more general system that could be used outside the LINKS environment. Hence, burgerLinker was developed in a collaboration between the IISH, Utrecht University and Vrije Universiteit Amsterdam (for a comparison of record linkage techniques, see [Christen, Vatsalan & Fu, 2014](#); [Raad et al., 2020](#)).

BurgerLinker is a graph-based record linkage program that uses the existing matching rules from the LINKS SQL-query environment, but retrieves candidate matches between certificates more efficiently. The program is designed as a stand-alone, scalable, and flexible software that allows the matching of other types of historical demographic records. Just as in LINKS, users can change the default settings for Levenshtein distance, ignoring filtering based on the dates of the certificates, avoiding matches without an identical first letter on the family name, ignoring parental names, or a mixture of the above. This leaves the desired precision and recall to the user, allowing users to see which candidate matches are filtered out and why. This contrasts with the MySQL environment, which was designed to deliver an optimal number of matches based on the discussions and decisions of the experts associated with the LINKS database. Users of previous releases could decide not to use certain matches which were flagged as weak, but were unable to create new matches based on their own criteria.

In the following, we will go into the processing aspect of the data and the data model, the working of the Levenshtein algorithm, the flexible recall and the increased transparency of the system.

4.3.2 BURGERLINKER PIPELINE AND DATA MODEL

BurgerLinker is graph-based and expects an HDT file (Header, Dictionary, Triples) as input, which is a standard format used within the Resource Description Framework (RDF). HDT compresses datasets in a significant way while maintaining efficient search and browse operations without prior decompression. The RDF format is a W3C standard (along with HTML, CSS, and XML) that models data through so-called triples, describing an entity with attributes and the value of attributes. In RDF-terminology, the entity is called the subject, the attribute is called the predicate, and the value is called the object. For example, "Nicholas de Vries" can be described as a person with the first name Nicholas. RDF writes this down in two triples:

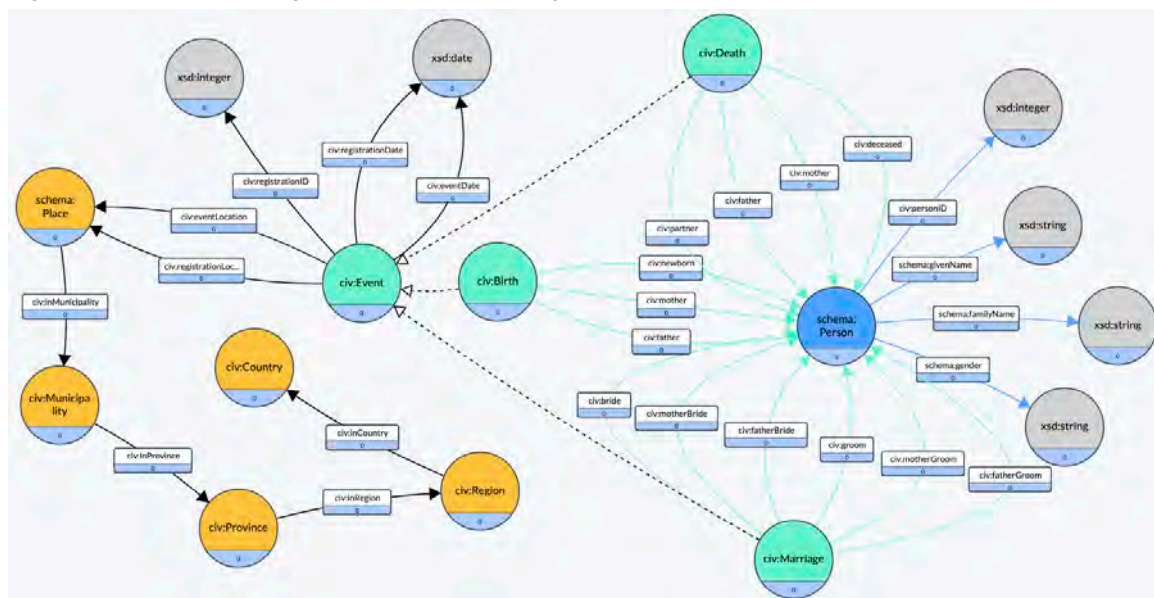
```
civ:personID_9999 rdf:type          schema:Person
civ:personID_9999 schema:givenName "Nicholas"
```

Each part of a triple is always represented by a so-called Uniform Resource Identifier (URI), which is similar to a Uniform Resource Locator (URL), except that they are global identifiers rather than only locators or addresses. The prepositions "civ", "rdf", and "schema" are abbreviations for so-called namespaces referring to existing vocabularies in which the content is defined. The local identifier and namespace ensure that all properties in the graph are not only unique in the LINKS environment but are also globally unique on the World Wide Web. In this example, "civ" refers to the IISH namespace, "rdf" to the w3c namespace, and "schema" to schema.org. The first triple says that the personID_9999, that was assigned a global identifier using the IISH namespace, is an instance of the class Person. This triple will be interpreted similarly across all RDF applications since it has been declared using the rdf:type property standardised by the W3C, and using the class Person defined by the schema.org vocabulary, commonly used in different applications and webpages. When using standardised vocabularies such as RDF and RDFS (RDF Schema), users can directly benefit from certain reasoning capabilities supported in most RDF platforms. For instance, given the following two triples in an RDF graph: civ:personID_9999 rdf:type civ:Male; civ:Male rdfs:subClassOf schema:Person. An RDFS reasoner allows us to infer on demand a third triple indicating that personID_9999 is also an instance of the class schema:Person, as the second triple declares that all instances of civ:Male are also instances of schema:Person.

Data for burgerLinker is exported from the LINKS MySQL database of standardized and cleaned data as a Comma Separated File (CSV). By way of the IISH standard tool COW, a script is run to convert the data from CSV to RDF.⁸ The required HDT input for burgerLinker is created by a) importing the data in RDF format and b) using an embedded tool to convert the RDF dataset into HDT. Within burgerLinker we can then match the civil certificates, and export the results as RDF or CSV.

⁸ The Python library COW (CSV On the Web; <https://csvw-converter.readthedocs.io/en/latest/>) allows flexible conversion of CSV datasets to RDF by relying on a JSON schema. For an example JSON schema that converts LINKS datasets from CSV to RDF according to the CIV model on the burgerLinker GitHub (https://github.com/CLARIAH/burgerLinker/blob/main/assets/examples/births_example.csv-metadata.json).

Figure 4 Civil Registries schema for burgerLinker



Explanation: BurgerLinker retrieves information on persons (blue), events (in green), and locations (in yellow) using the schemas Civil Registries schema (civ) and schema.org (schema). The gray indicates the expected type of literal values. Extra attributes can be added to the model, for instance, occupation, address, or age for persons, the name of the clerk for registrations, or the names of witnesses for events. Note that some of these variables are already defined in the Civil Registries Schema, such as age or occupation. See also Appendix B.

BurgerLinker requires that the data are modelled according to a so-called schema. A schema consists of classes (and instances of classes), which are the main entities of the data structure. The classes are associated with (sub) schemas describing specific cases of persons or locations. The schema designed for civil certificates is included in Figure 4 and is named "Civil Registries schema" (CIV). Figure 4 describes the core parts of this data model. The main entities or classes are presented as nodes. We see four (green) nodes for events: three classes for each type of civil certificate: birth, marriage and death and a more general class for the information about the event itself, heading schemas for dates and places both for the event itself and the registration of the event. Each event has a defined set of persons, for example, child, father and mother in the birth certificate of which the attributes are defined in the schema Person. The schemas are defined in our own CIV model or derived from existing schemas, in this case "xsd" for the data type (date, integer or string) and "place" for locations linked with our own more general schemas for the municipality, province, region and country. Each arrow in Figure 4 represents a triple pattern which defines the roles of the persons involved in an event and the attributes that are included in the model, for example, the name with the gray node indicating the type of literal values that are expected. In Appendix B we have included a more formal and complete description of the CIV model.

After matching, the results need to be combined with other retrieved data such as professional titles and combined into event histories and family reconstitutions (Mourits et al., 2020).

4.3.3 OPTIMIZED LEVENSHTEIN ALGORITHM

Just like the SQL-query environment, burgerLinker uses a Levenshtein algorithm to match cases. Calculating Levenshtein distances is a time-costly process, as the number of possible matches that a Levenshtein algorithm needs to consider grows exponentially with each new name that is added to the database, thus increasing the required run-time exponentially. In the MySQL database this problem was solved by making prematch tables that store the Levenshtein distance between unique names before starting the actual matching procedure. These prematch tables made the linking within the SQL-query environment more efficient, but are an extra step in the linking process that must be reproduced when new data are added to the system. In burgerLinker, the Levenshtein distances are calculated efficiently on the fly. To speed up the computation of Levenshtein distances, burgerLinker indexes the list of target names as a Minimal Acyclic Finite-State Automaton (MA-FSA), also known as Directed Acyclic Word Graphs (DAWG). An FSA is a mathematical model or an abstract machine that operates by moving

through a series of states in response to inputs, where each state represents a particular condition or configuration of the machine. Then, a Levenshtein transducer is initialized, which is an FSA that accepts a query term (e.g., a name) and returns all terms in the index that are within n spelling errors away from it. Like the MySQL system, burgerLinker allows the user to specify the maximum accepted distance for a match, with 4 being the maximum allowed distance in the current version. This procedure, implemented in the JAVA library *liblevenshtein* based on the work of Schulz and Mihov (2002), is much more efficient than the original Levenshtein algorithm, as its runtime complexity grows linearly with the length of the query term, rather than exponentially on the size of the index (Raad et al., 2020).

4.3.4 FLEXIBLE RECALL AND INCREASED TRANSPARENCY

In its earliest stage, burgerLinker produced the same matches as the SQL-query environment, using the same matching principles. However, during the testing of burgerLinker we changed strategies and opted to aim at retrieving as many candidate matches as possible. The policy for the construction of releases in the SQL environment focused on finding as many unique matches as possible within the Dutch civil registry, prioritizing the quality of established matches and limiting the retrieval of candidate matches as soon as too many multiple matches appeared. This restrictiveness on the number of matches was advantageous for researchers as the retrieved dataset was ready for analysis. Yet, this optimum is not always the same for different datasets and there is also some variance between disciplines in what researchers deem the optimal balance between recall and precision (see Section 5.2 for an elaborate discussion of this balance). The structure and matching speed of burgerLinker make it relatively easy to match with different alternative designs and define how we filter matches to get more precision at a later stage.

Secondly, by increasing the importance of the filtering procedure after the matching, burgerLinker makes the whole matching process more flexible and transparent. Just like the LINKS query system of the MySQL database, burgerLinker provides extensive data on the background of these matches. However, since burgerLinker can be used as a stand-alone tool, users are independent from the database manager in running the matching program, and can decide the maximum Levenshtein distance per name on the spot, as well as the number of persons on a certificate that should match.⁹ Just as in the SQL-environment, the Levenshtein distances are made dependent on the character length of the smallest item to be matched (see Table 5 and compare Table 4). By giving users the possibility to retrieve a larger set of candidate matches, the matching procedure becomes more transparent, as it becomes clearer which candidate matches are rejected to get a higher precision.

Table 5 *Changeable settings in the burgerLinker matching environment*

Settings	Default
Maximum Levenshtein distance	4
Fixed Levenshtein distance	False
Ignore date consistency check	False
Ignore blocking first letter last name	False
Match single individual	False

Max Lev Distance	Restriction of Lev Distance based on name length				
	0	1	2	3	4
0	1+	-	-	-	-
1	1–5	6+	-	-	-
2	1	2–8	9+	-	-
3	1	2–5	6–11	12+	-
4	1	2–5	6–8	9–11	12+

Explanation: A choice for a specific maximum Levenshtein distance automatically sets lower values in case of relatively short name lengths. So, Levenshtein 4 will not work for names smaller than 12 characters.

⁹ See <https://www.github.com/clariah/burgerlinker>

The way in which first names are matched was slightly altered to retrieve more candidate matches. Just as in the LINKS system, each first name is matched separately, rather than considering them as one string. For example, "Hendrikus Kornelis Romein" can now match to "Hendrikus Romein" and "Kornelis Romein." The difference is that in burgerLinker no choices have to be made in deciding how and which part of the multiple first names will be matched (compare Section 4.2.2, Table 4). Although this procedure can lead to some overmatching, it also reveals many potentially useful matches. However, to limit obvious mismatches, two entries with multiple first names will not match when separate elements of the shortest multiple first name does not match. For example, "Hendrikus Kornelis" can match to "Hendrik Kornelis Aloysius", but not to "Johannes Hendrikus Wilhelm", because the last one lacks "Kornelis." Systematic filtering is possible due to the detailed metadata on the number of matched first names, the Levenshtein distance per matched first name, and the total Levenshtein distance of all matched first names. As a result, we can still retrieve the same matches as the SQL-environment, provided that we use the same rules for filtering.

Like the database manager in the LINKS MySQL-system, users of the burgerLinker system may choose to ignore the date consistency check, ignore blocking of the first letter of the last name, or match only on the name of one of the indexed persons (see Table 5). In general, matching on one person is not advisable, since it will give an enormous number of false matches. By default, the system will match on the ego and the mother and additionally on the father and a partner if possible. We also made it easy for the user to decide if children should be born in wedlock or not (see discussion in Section 3.2).

The main difference between burgerLinker and the MySQL query environment is that burgerLinker is designed as a tool for general use. Our hope is that burgerLinker can serve as a tool to make matching procedures within historical demography easier for researchers and also more comparable by introducing a standard way of matching. The introduction of the Intermediate Data Structure (IDS) for life course databases (Alter & Mandemakers, 2014) and its wide acceptance in the field, have laid the basis for common software, but the system supposes that the record linkage is done by the database itself. Because each database has its own selection of sources with their own local peculiarities, database managers have developed a wide range of different matching strategies. BurgerLinker will not replace these existing matching programs, but it can easily link new data to existing datasets. It could also be used in validating the quality of existing matches and help to make alternative matchings which deviate on some aspects from the standard release. This will be facilitated by converting unlinked data into IDS. The IDS is structured according to the principle of the Entity Attribute Value model (EAV) or object-attribute-value model, which was introduced in the 1970s (Stead, Hammond, & Straube, 1982). This is exactly the same structure as the RDF triple system in which the subject is the entity, the predicate is the attribute, and the object is the value (<https://graphdb.ontotext.com/documentation/9.8/enterprise/devhub/rdfs.html#what-is-rdf>). This makes the creation of conversion scripts to create triple systems a relatively easy job. Hence, burgerLinker is the HSNDB's effort to share our experience in matching civil records with the community.

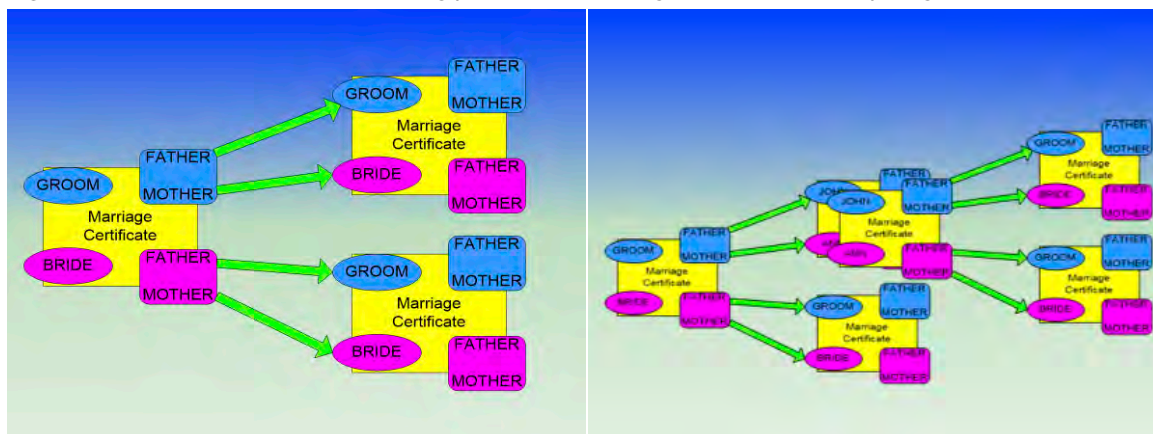
5 MATCHING EXPERIMENTS WITH MARRIAGE CERTIFICATES

5.1 THE CASE OF ZEELAND

In this section we will show the result of our experiments with the matching system, limiting our evaluation to the marriage certificates of the province of Zeeland. We choose the province of Zeeland because of the completeness of the dataset and the relatively limited number of inhabitants (on average about 5% of Dutch population).

Figure 5 presents the challenge: the father and mother of a bride or a groom are to be matched with their own marriage certificate. As soon as a match is found we have a family tie between three generations. Since both bride and groom are matchable with the marriage of their parents, we work along two lines which we call the bride and the groom line. After matching we may combine the matched pairs of certificates into lines of multiple generations. This kind of matching is relatively easy and straightforward since the *identifying* information of the marriage certificate of the second part of a pair will be identical with the first part of another pair. Essentially, this is a process of matching pairs of persons to get linked certificates.

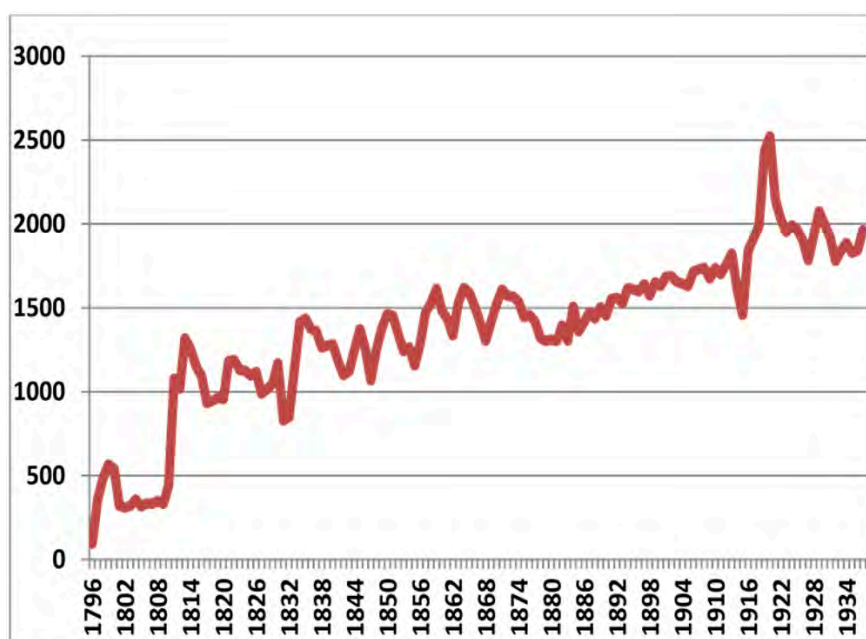
Figure 5 Scheme of the linking process of marriage certificates into pedigrees



For Zeeland there are indexed certificates for the period 1796–1936. See Figure 6 for an overview of the numbers per year. For most of the country, civil certificates were introduced in 1811. However, the southern part of Zeeland (Zeeuws-Vlaanderen) had started recording certificates in 1796, because they were annexed by France in 1795 (Vulmsa, 1988). The rise in the numbers in 1811 is explained by the inclusion of the rest of the province Zeeland. The strong fluctuations between 1916 and 1920 were a consequence of the bad economic situation during the last two years of the First World War (although the Netherlands were not involved in the fighting) and the subsequent optimism in the three years after the end of this war (van der Bie, 1995; van Zanden & Griffiths, 1989).

All in all, the total number of included marriage certificates amounts to 191,847. This number excludes certificates before 1801, because we found a lot of double registrations of the same marriage in different municipalities, which would systematically result in double matches. Since each marriage consists of two partners, we can expect theoretically twice the number of matched parental certificates: 383,694. However, in practice not all certificates could be matched with those of the parents due to two main restrictions. The first one concerns the time window of observation: parents of persons who married before 1830 will have no chance to be matched with their own marriage certificate, since the civil registration started only in July 1811 (with the exception of some from Zeeuws-Vlaanderen). Similarly, the parents of persons who married between 1830 and 1860 are estimated to have on average only a 50% chance to be matched to their own marriage certificate. If we take this into account, there are only 339,240 potential matches left. But this number will not be reached, because — and this is the second restriction — an unknown proportion of parents who married outside Zeeland were not included in the matching operation.

Figure 6 Number of marriage certificates per year, Zeeland, 1796–1937



5.2 RESULTS

To measure the quality of our results, we estimate recall and precision. In record linkage the term "recall" stands for the relative share of retrieved matches compared to all potential matches, and a higher recall indicates fewer missed matches, also called false negatives. Precision is the relative share of correct matches compared to all matches made, and a higher precision indicates fewer false matches, also known as false positives. False positives are most clear when more than one match is found for the same record. This kind of multiple matching we call overlinking. In our case, overlinking exists when the parents of the groom or the bride are matched with more than one other certificate, which can only be true in the very rare case that parents remarried after having had a divorce.

The ideal in record linkage is to arrive at zero false negatives (recall) and zero false positives (precision). Since there is no information on the actual number of correct matches, we need to use other indicators to get an impression of the quality of the matching. Therefore, we use the total number of retrieved matches as an assessment of recall and the number of multiple links for the same certificate as an assessment of precision. This procedure allows us to test different matching criteria, where we go from very strict matching criteria to more free ones, while we stop when the number of matched certificates increases marginally and the number of multiple links expands exponentially (Oosten, 2008).

We combined three different methods of matching. Firstly, name matching varies depending on blocking on the first character or not in combination with the setting of the Levenshtein values. Secondly, we matched multiple first names in three variants: a) the two first names combined, b) only the first name and c) one name out of all names. And thirdly it is possible to remove checking the date range in which a parental marriage should be expected. The results of our matching operation or shown in Table 6.

We start with a baseline of very strict matching in which we a) match both for family names and first names exactly with Levenshtein=0, b) include the two first parts of a first name (if present), and c) accept no matches that are outside the estimated marriage range. This is the strictest matching set-up the system offers. The right half of Table 6 shows the result of each matching method in a) terms of the number of matches and overlinks and b) by way of an index that show the relative differences with the baseline. The right-most column shows the ratio between both indexes. If the ratio stays close to one, it indicates that the matching operation does not produce relatively more overlinking than the more secure matching operations.

The baseline shows a result of 163,237 close to matches. That is almost half of the theoretical total ($n=339,240$). The number of overlinks is 158 which is less than 1 in 1000. The names on the overlinked certificates are identical and all fit the time range, meaning that they cannot be distinguished by the matching software. In a second step, we raised the Levenshtein values to 2. This resulted in 10.4% extra matches and 11.4% extra overlinks, which corresponds to a ratio of 1.01. In other words, it does not make much difference when Levenshtein 0, 1 or 2 is used in the matching procedure, the quality of the matching remains the same and we have 16,955 extra matched certificates. Choosing Levenshtein 4 did not show much difference either; compared with Levenshtein 2 it returned 11.0% extra matches and 18.4% extra overlinks, which corresponds with a ratio of 1.07.

In the second group of matching exercises, we experimented with 'freeing' the first character. This was only done for a maximum Levenshtein value of 2, since earlier experiments with unblocking the first character and accepting higher Levenshtein values resulted in an explosion of false matches. We see that freeing the first character of the first name with Levenshtein 2 does add 2,381 matches (1.3%) compared to the fixed variant with a ratio of 1.02 (compared to 1.01). The other options which include freeing the first character of the family name or accepting a Levenshtein value of 4 for the first name results in ratios between 1.09 and 1.14 which do not differ much from each other nor from the baseline. And in the last option we have 20,377 (12.5%) more matches than from the baseline settings.

In the third group, we changed the way the first name is handled. At the beginning of our period, only 30% of the persons born in Zeeland had two or more names, at the end, this percentage had risen to almost 60%, of which 10% consisted of three or more names (Gerritzen, 1998). So, for roughly half our population, there might be different results depending on how the first name is handled. Restricting the first name to the first part gives an extra 9,546 matches compared with the baseline (Levenshtein 2, fixed first character, two name parts) and 26,501 more compared with the baseline with the exact match. The overlinking increases within reasonable limits from 158 to 232, which corresponds with a ratio of 1.26. The freer variants in this group result in more matches with a relatively low number of overlinks (ratio 1.33 and 1.53).

Table 6 Results of the matching process, Zeeland, marriage certificates, 1801–1937

Familyname		First name		Dates	Quality indicators					
First character	Maximum Levenshtein	First character	Maximum Levenshtein	Elements used	Ranges	Matches		Overlinks		Ratio
						N	Index	N	Index	
fixed	0	fixed	0	1+2	fixed	163,237	100.0	158	100.0	1.00
fixed	2	fixed	2	1+2	fixed	180,192	110.4	176	111.4	1.01
fixed	4	fixed	4	1+2	fixed	181,114	111.0	187	118.4	1.07
fixed	2	free	2	1+2	fixed	182,573	111.8	181	114.6	1.02
free	2	fixed	2	1+2	fixed	181,323	111.1	195	123.4	1.11
fixed	4	free	2	1+2	fixed	182,826	112.0	193	122.2	1.09
free	2	fixed	4	1+2	fixed	182,002	111.5	197	124.7	1.12
free	2	free	2	1+2	fixed	183,714	112.5	203	128.5	1.14
fixed	2	fixed	2	1	fixed	189,738	116.2	232	146.8	1.26
fixed	4	fixed	4	1	fixed	190,548	116.7	245	155.1	1.33
free	2	free	2	1	fixed	193,295	118.4	287	181.6	1.53
fixed	2	fixed	2	1 of all	fixed	196,062	120.1	320	202.5	1.69
fixed	4	fixed	4	1 of all	fixed	196,951	120.7	336	212.7	1.76
free	2	free	2	1 of all	fixed	199,551	122.2	411	260.1	2.13
fixed	2	fixed	2	1+2	free	183,276	112.3	559	353.8	3.15
fixed	2	fixed	2	1 of all	free	200,595	122.9	1,520	962.0	7.83
fixed	4	fixed	4	1 of all	free	201,575	123.5	1,589	1005.7	8.14

Explanation: The columns "First character" indicates if the first character is fixed (or blocked) in the matching process of free. The column "Elements used" indicates how a composed first name is matched ("1" only the first one, "1+2" only the first two ones and "1 of all" only one random element with another one). The column "Dates" indicates if the estimated ranges within parents will marry are used to limit matching possibilities. Free means that there was no check on these ranges.

In the fourth group, we again changed the way the first name is handled. We matched in such a way that each part of the first name had an equal chance to be part of the match. A first name like "Maria Elisabeth Antonia" will match "Maria", "Elisabeth" and "Antonia." The other settings are the same as in the third group. If we compare the results, we see that for all three lines this action results in about 6,300 extra matches. Looking at the maximum number of matches compared with the baseline with exact matches, there are 36,314 extra matches while still having only a moderate increase in overlinking to a ratio of 2.13.

In the fifth and last group of Table 6, we removed the constraints on the minimum and maximum ranges of parental marriage. We see that the number of overlinks increases with a little gain in matched certificates. This results in ratios that vary from 3.15 to 8.14. In the first case, which blocked only the first character of the first name, there are 3,084 matches more than the comparable matching with fixed date ranges; while the number of overlinks increases threefold from 176 to 559. In the last two rows, we matched all parts of the first name separately. Although we got respectively 4,533 and 4,624 extra matches, compared with the first row where we matched one first name to all other ones ("1 of all"), the number of overlinks grew almost fivefold, resulting in a ratio of respectively 7.83 and 8.14 against the exact baseline. Given this last result we conclude that 'freeing' logical date ranges is a bad strategy, unless it is done in a very limited way (compare Section 3.2).

On the basis of this experiment, we concluded that in the case of the marriage certificates, it did make a small but not unimportant difference to free the first character (given a Levenshtein level of 2). In practice, we found that most of the cases were typical features of Dutch language: mixing up of "y" and "ij", "c" and "k", "f" and "ph", "s" and "z", "ch" and "g". Because freeing the first character is very expensive in terms of computing time, we decided to stay with fixing the first letter, but standardizing the family names as described in Section 3.1.

All in all, we consider the result of our experiments as a proof of the excellent quality of the data in general, especially comparing our results with linkage between American censuses (Goeken, Huynh, Lynch, & Vick, 2011). The difference between the result of the exact matching and the most flexible alternative was only 38,338 matches (23.5%). Two factors contribute to our success. First, the high quality of the data which was due to legal requirements governing the civil registration system, especially the obligation to submit official extracts of birth certificates as part of marriage registration (Vulsma, 1988). Second, since females retained their own family name, we are usually matching pairs of people instead of individuals.

What matching strategy can be distilled from our experiments? We learned that standardization of typical Dutch spelling variances limits the advantages of freer matching, especially above the limit of Levenshtein 2. Secondly, matching with date ranges is very useful in limiting overlinking. That leaves the question of how to match with multiple first names. Should we use only the first name or the first two? Alternatively, can we use an algorithm in which one part of a first name will be sufficient, independent of its place in the sequence of names? For Zeeland, matching the first name compared with matching the first two names, resulted in more matches with a limited increase in overlinking. Comparing one part of a first name with all other parts, also seems acceptable. However, the degree of overlinking probably will be higher in other parts of the Netherlands as about half of the Zeeland population had only one first name. We may expect that the degree of overlinking grows exponentially when the entire population has multiple first names, especially higher social groups and the Roman Catholic part of the population which was very generous in giving multiple first names (Bloothoof & Onland, 2016; Gerritzen, 1998). On the other hand, less precision can be acceptable when all types of certificates are linked to produce family reconstitutions that can be used to test the integrity of the whole family (see Bloothoof, van Boheemen, & Schraagen, 2016; van Boheemen, 2016).

We can conclude from the results in Table 6 that about 199,000 matches is the maximum that we may expect from this dataset. We calculated a theoretical total of 339,000 matches. This implies that about 140,000 (42%) of the parents married outside Zeeland during the period from ca. 1810 till 1910. We may test this in future 1) by linking the certificates of Zeeland with certificates covering the rest of the Netherlands and 2) by adding matching algorithms working with root names. We expect more matches using roots for the first name, because they will take into account abbreviations, such as Jan instead of the "Johannes", and translations, such as "Guillaume" instead of "Willem" (Oosten, 2008). We will also select cases to be examined manually to find software bugs and dataset errors, such as certificates that have been entered twice in the original dataset, and to determine if some double links are really couples who married each other for a second time. Future versions of LINKS will keep using the number of matches and overlinks as proxies for recall and precision. The optimal settings will probably differ by region and time period, but the ratio will help determine the optimal settings for each context.

6 RELEASES

6.1 INTRODUCTION

In the foregoing, we explained how LINKS operates in cleaning and matching the civil certificates from the WieWasWie indices. However, matched certificates are only the basis for a research dataset. As this is being written, 40 datasets have been constructed, of which five can be downloaded directly (<https://datasets.iisg.amsterdam/dataverse/hsndb-links>). For a full overview of all releases, see <https://iisg.amsterdam/en/hsn/projects/links/links-releases>. Most of these releases covered only parts of the country, dependent on the availability of indices, the transcription of occupational titles and specific requests from researchers. Over half of these releases were 'forerunners', to be used by researchers for testing the quality of the data or developing the program to construct the dataset for analysis or doing preliminary analyses. Until 2010 only the marriage certificates were indexed completely enough to link them and to make meaningful releases (of pedigrees). It took more time to index the other certificates, and the birth certificates were lagging behind. The first more or less completely indexed provinces were Zeeland, Limburg and Groningen/Drenthe. After 2010 the index improved enormously in quantity (see Table 1) and it became possible to link the country as a whole

which resulted in two large releases at the national level: all marriage certificates linked into pedigrees (Mandemakers & Laan, 2020b) and all birth certificates linked with the marriage certificates of the parents (Mandemakers & Laan, 2020a).

Ultimately, the main goal of LINKS is to deliver complete *integrated* datasets of births, marriages and deaths, creating families and multigenerational links. So far, this kind of dataset is only realized for the provinces of Zeeland, Limburg and Groningen/Drenthe separately (Mandemakers & Laan, 2017, 2018, 2019). Researchers have used these datasets to create two types of datasets suitable for statistical analysis. The first one was a rectangular data structure constructed within the context of the project Genes, Germs and Resources (Mourits et al., 2020). The second one was a reconstruction of the Zeeland release in the format of the Intermediate Data Structure (IDS; Alter & Mandemakers, 2014).

In the following we will explain first the construction and quality of the national releases and secondly the integrated ones.

6.2 THE NATIONAL RELEASES

6.2.1 MARRIAGE CERTIFICATES

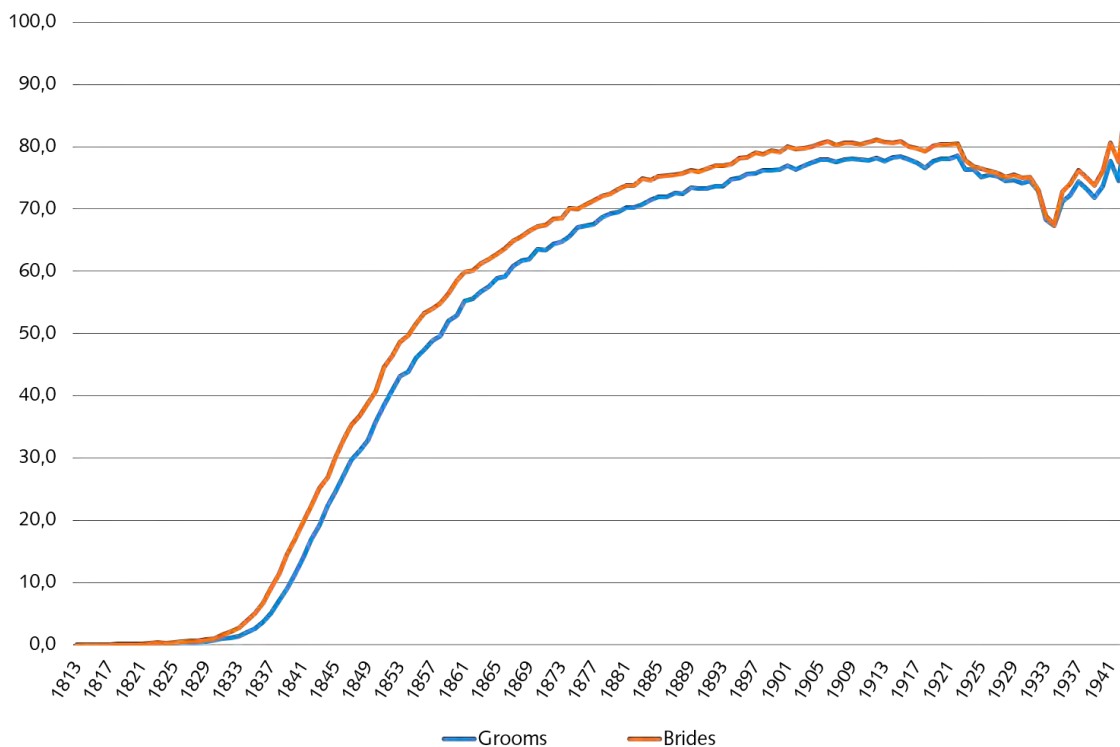
The release of the marriage certificates and their parental links (Mandemakers & Laan, 2020a) included a total of 4,158,387 certificates. The marriage certificates were linked into pedigrees (see Section 5.1 for this process). Actually, we handled 8,316,774 "marriage lines", one for the parents of the bride and one for those of the groom. The matching procedure was based on the Zeeland experiments (see Section 5.2) and consisted of four elements:

- 1 First names and family names were standardized as described in Section 3;
- 2 The first name and family name of each person were separately matched with an accepted level of variance of maximal Levenshtein 2;
- 3 If the first name of a person consisted of more than one part only the first part was used for matching, so a first name as "Cornelis Albert Maria" was restricted to "Cornelis";
- 4 The date of the parental marriage has to be 14 to 49 years before the date of the marriage certificate in which they show up as parents. This range was based on the childbearing ages of the mother and was further limited if the age of the newlyweds was indexed as well.

Matching pairs of persons means that four different strings were compared and matched: two first names and two family names which implies that Levenshtein distances could be as high as 4×2 equals 8. Some persons married more than once, which is indicated by the civil status of the bride or groom. However, this kind of information is not included in the index in a systematic way. A second matching between the parents of the marriage certificates themselves (see Section 6.3.3) could provide this information, but this operation has not been done on the national level yet. The indexed information from the certificates is limited. As mentioned in Section 2, all archives include at least the municipality and date of the event as well as the first name and family name of the bride, groom and their parents and usually the age at the event for the bride and groom. Occupational titles have been transcribed for about 60% of the marriage certificates covering seven provinces (out of a total of eleven, see Figure 1).

Of the indexed marriage certificates, 99.3% date from the period 1812–1943 when civil registration was obligatory for the whole of the Netherlands. Most of the indexed certificates are from before 1922, since indexing is lagging behind the public release of certificates. Figure 7 shows the percentage linked to the parental marriage by year of marriage. This is almost zero before 1830, because parental certificates seldom show up before the age of 18 (of the parents). From 1830 until 1860 it increased to about 60% and then further rose to about 80% in 1920. The fall and the rise after 1920 are a result of the uneven development of the index. Brides always do slightly better than grooms, because marriages tended to take place in the birthplace of the bride, resulting in a better chance to link the parental marriage certificate.

Figure 7 *Relative number of linked certificates per marriage line, Netherlands, 1812–1941*



Source: LINKS dataset linked marriages (Mandemakers & Laan, 2020b).

Table 7 presents an overview of the matching results. In total, 59.8% of the brides and grooms were matched with a parental certificate. Parental marriages before 1812, mostly explain why 28.1% of the certificates were not matched. The remaining 12.0% did not link for other reasons, such as lacking indices, ambiguous matching results or foreign marriages. Brides or grooms linking with more than one parental certificate are marked in the dataset and not linked. An exception are cases with two alternatives of which one match was almost exact and the other one had a relatively high score on Levenshtein. On this basis a meagre 0.1% could be added to the linked results. These decisions are marked in the release tables, so a user may decide not to use these ambiguous links.

Table 7 *Number of marriages lines and matching results with parental marriages*

	Number	Percentage of total
Link with parental certificate	4,975,177	59.8
No ambiguous link	4,964,157	59.7
Ambiguous but reasonable choice	11,020	0.1
No link because of technical reasons	2,337,205	28.1
Ambiguous linking result (two or more links)	217,072	2.6
Lacking identifying data of one parent	99,557	1.2
Lacking identifying data of both parents	372,206	4.5
Time range (marriage certificate before 1830)	724,783	8.7
Time range (marriage certificate 1830–1860, estimation)	923,587	11.1
No link because of other reasons	1,001,590	12.0
Total	8,313,972	100

Source: LINKS dataset linked marriages (Mandemakers & Laan, 2020b).

Table 8 *Total Levenshtein value of established links marriages with parental marriages*

	Number	Percentage of total
Exact match	3,948,798	97.4
Total Levenshtein value = 1	687,795	13.8
Total Levenshtein value = 2–3	315,738	6.3
Total Levenshtein value = 4–8	22,846	0.5
Total	4,975,177	100.0

Source: LINKS dataset linked marriages (Mandemakers & Laan, 2020b).

The sum of Levenshtein distances is also made explicit for each matched certificate (see Table 8). Of all matches 79.4% proved to be an exact match and only 0.5% were matched with a total Levenshtein value of three and more. This clearly indicates the good quality of the Dutch certificates and the matching operation. That 2.6% of the certificates with more than one match remained to unresolved (as shown in Table 7) is because most of these matches are of very good quality in terms of Levenshtein distance.

6.2.2 BIRTH TO PARENTAL MARRIAGE LINKAGE

Another release on a national scale is the linkage of births with parental marriages (Mandemakers & Laan, 2020a). This kind of matching implies that both parents must be known on the birth certificate, excluding all illegitimate children. In the future, some of these births may be linked in an indirect way (through linking the child and its mother in the death certificate or marriage certificate of the child).

We used the same matching conditions used for the marriage certificates (Section 6.2.1). We also formulated a comparable time range on the basis of the age of the bride at the event of the marriage. This means that the child's birth date should be between the date of marriage of the parents and the date of marriage plus 49 years minus either the age of the bride at marriage or 14.

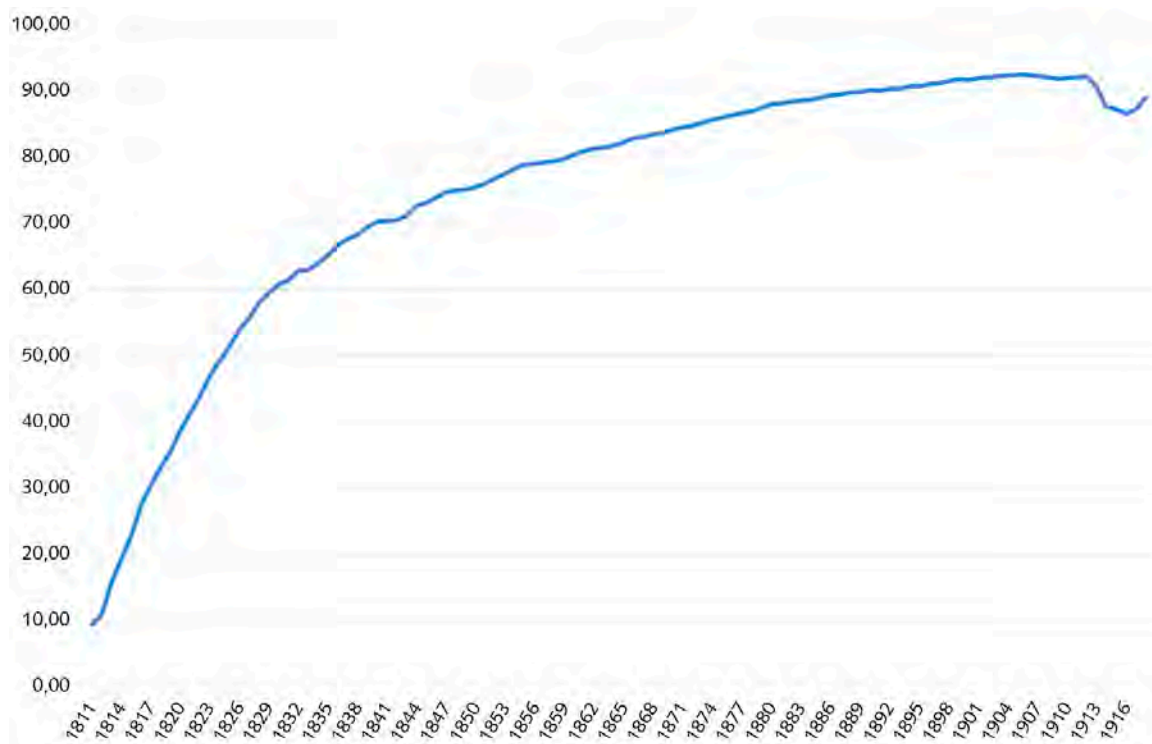
The index on birth certificates is less complete than the one for marriage certificates, e.g., Amsterdam is lacking completely. Other archives were only partially covered or had indexed only the name of the birth and no parents. To make a consistent release, we included only births from archives which had a matching rate of at least 60%, totaling 9,792,024 birth certificates which are about 2/3 of the potential number of births (compare Table 1). Of these birth certificates, 7,669,986 were linked with at least one marriage certificate (see Table 9). For 21.7% of the births no link was found. The main reasons for missing links are parental marriages before 1811 (11.3%) and incomplete data about the father (births outside a wedlock, 1.5%). For 8.5% of the missing links there is no clear reason, but some marriages are not included in the marriage index yet, and some marriages were registered outside the Netherlands. In the release we identified 25,020 births with more than one linked marriage certificate that could not be resolved. For 16,704 cases we made a choice for a specific certificate in a comparable way as we did with the marriage certificates (see Section 6.2.1).

Table 9 *Number of births and matching results with parental marriages*

	Number	Percentage of total
Link with parental certificate	7,669,986	78.3
No ambiguous link	7,653,282	78.2
Ambiguous but reasonable choice	16,704	0.2
No link because of technical reasons	1,289,625	13.2
Ambiguous linking result (two or more links)	25,020	0.3
Lacking identifying data of father	148,157	1.5
Lacking identifying data of mother	7,984	0.1
Time range (marriage could be before 1812)	1,108,464	11.3
No link because of other reasons	832,413	8.5
Total	9,792,024	100.0

Source: LINKS dataset linked births and parental marriages (Mandemakers & Laan, 2020a).

Figure 8 *Relative number of births linked with parental marriages, 1811–1918*



Source: LINKS dataset linked births and parental marriages (Mandemakers & Laan, 2020a).

Figure 8 shows the share of matched birth certificates per year. After 1850 one could expect that each birth with two parents will match with a marriage certificate. This is not always the case for reasons already mentioned. Around 1850 the percentage is about 75%, climbing to 92% for the period 1898–1912, dropping in the years of the First World War to 87%.

Looking at the Levenshtein distances, we found more or less the same results as presented in Table 8 for the linking of the marriage certificates. Of all matches, 79.6% proved to be an exact match and only 0.4% matched on the basis of a total Levenshtein value of three and more. This is another clear indication of the good quality of the Dutch certificates.

6.3 THE REGIONAL INTEGRATED RELEASES

Given the state of the indices at the end of 2017, it was possible to create integrated sets of birth, death and marriage certificates for four provinces: Zeeland, Limburg and the combination of the two bordering provinces Groningen and Drenthe (see Figure 1). For reasons of research the indices needed the inclusion of occupational titles, ruling out provinces as Utrecht and Friesland which also have high levels of indexing.

In the following we explain how we linked the different certificates, how we created uniquely identified persons out of these data and how we changed the pedigree structure into a more family tree-like structure. Firstly, we will describe the nucleus of the table system that was created out of the linked certificates. Secondly, we will elaborate on the linking process, going into the several types of matching that needed to be made. Thirdly, we will explain how the persons in all the certificates were synchronized into unique persons. In the last section we will elaborate on the outcome in terms of unique persons and families.

6.3.1 STRUCTURE OF THE DATASET

The resulting dataset consists of a system of four interlinked tables. See Figure 9 for the table structure and the relationships between the tables and the identifying keys.¹⁰

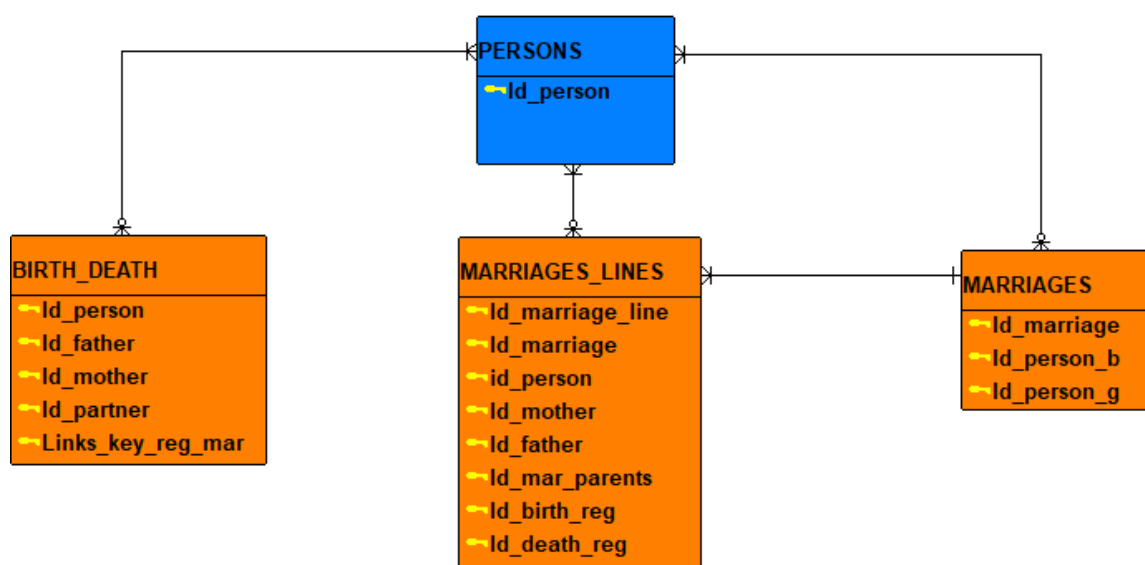
10 For the sake of clarity, the names of some keys have been changed in comparison with the published documentation (Links_key_reg_mar in BIRTH_DEATH became Id_mar_parents and Id_mar_parents in MARRIAGE_LINES became Id_mar_parents_intern).

The table BIRTH_DEATH establishes the link between a birth and a death certificate and includes all information from these certificates, such as place and time of death, occupational titles, etc. The table includes not only the linked but also the unlinked birth and death certificates. The key Id_partner refers to the (last) spouse in the death certificate.

The data of the marriage certificates is included in two tables: MARRIAGES and MARRIAGES_LINES. MARRIAGES includes all information about the marriage itself, including the identifiers of the bride and the groom. MARRIAGES_LINES contains the personal data of the bride or the groom and the data about their parents. The key Id_Marriage identifies each marriage. So, each marriage produces one record in MARRIAGES and two (bride and groom) in MARRIAGE_LINES (see also Section 6.2.1). The link with the marriage of the parents is defined through the (internal) key Id_mar_parents_intern which refers to Id_marriage.

Births and deaths belong to families, this relationship is represented through the key Id_mar_parents referring to Id_Marriage in the table MARRIAGE_LINES [and the table MARRIAGES]. Each person may marry not at all or once or more, having the corresponding number of records in MARRIAGE_LINES linked by way of the key Id_person.

Figure 9 Table structure of interlinked civil certificates and unified personal information



All unique persons are included in the table PERSONS, which was constructed by including appearances of persons from certificates in the following sequence:

- 1 All persons (birth, mother and father) from the birth certificates;
- 2 All persons (death, mother and father) from the death certificates that were *not* linked with a birth certificate;
- 3 The last partner from the death certificates;
- 4 All persons from the marriage certificates (bride, groom, mothers and fathers) that are not known (= are not linked) from the birth and/or death certificates.

All persons included in PERSONS are linked with the underlying tables in Figure 9 with the identifiers id_person, id_mother, id_father, id_partner, id_person_b and id_person_g.

The first three steps are rather straightforward. However, adding of the marriages is more complicated, and an update of the identifying keys in BIRTH_DEATH is needed after adding the marriage certificates. So, it is not a question of simply adding persons. In the following section we will explain the construction of unique persons out of all their appearances in the several certificates.

6.3.2 IDENTIFICATION PROCESS OF UNIQUE PERSONS

For a complete construction of unique persons, we need five types of links:

a) Marriages and parental marriages (pedigrees)

The linking process contained only one step: The marriage certificates were linked on the basis of a link between pairs: the parents of a bride or a groom with a bride/groom couple, see Section 6.2.1 for the details of this matching. Links could not be established in case the parental marriage originates from the period before 1812.

b) Shadow marriages

In case a birth certificate was not linked with a parental marriage certificate, 'shadow marriages' were created. This matching works more or less in the same way as the one creating pedigrees. But here, the parents mentioned in different birth certificates are linked to form parental environments. So, the linking is based on pairs of two persons and these marriages are also bound within an acceptable time range. Shadow marriages of parents may go back far into the 18th century.

c) Births and Deaths

The linking process connecting the birth and death certificates, forming basic lifelines, contains two different approaches: a) linking on the basis of three persons: child, mother and father and b) linking of two persons: child and mother. The second option principally implies that no father is known. Of course, there will be cases of linked certificates in which fathers show up, who did not match in the first approach. We did not use this information because including these fathers could conflict with positively matched fathers from marriage certificates. Since the data about fathers is included in BIRTH_DEATH, a user can determine whether a father is known or unknown.

d) Births/Deaths and the Marriages of the parents

The link of a birth or death with the marriage certificates of their parents was created from the point of view of birth and death. First, we tried to match each birth to a marriage certificate. Next, we repeated the same operation on all deaths that were not linked with a birth certificate, including infants recorded in the death register but not the birth register, most of whom died shortly after birth. This procedure implies that in case the link from the birth certificate would provide different results than the death certificate, the former was given automatic priority. This choice was based on the legal requirement that brides- and grooms-to be had to show a birth certificate before a marriage could take place. This means that birth certificates were used to fill in the personal information on a marriage certificate. All matching was based on linking these two pairs: bride & groom and mother & father. In this way we created families, so a family is defined as all persons linked with the same parental marriage certificate. This implies that persons whose birth and death certificates were not linked could appear as siblings in the family tree.

e) Births/Deaths and Marriages

The link of a birth certificate with his or her own marriage certificate was made from the point of view of the bride and groom lines (the table MARRIAGE_LINES). This was done, because a person is only born once, but may marry more than once. The linking proceeded in two steps: a) Linking on the basis of three persons: child (bride or groom), mother and father, b) linking on the basis of two persons: child (bride or groom) and the mother to include also brides and grooms born out of wedlock. The linking of a death certificate with marriages is comparable, except that here two additional approaches can be used: matching on the basis of four persons: the deceased, mother, father and partner, and matching on the basis of the deceased and his/her partner.

After the linkage of the births and deaths with their marriage certificates, it was possible to add more links between the birth and death certificates. Parents were often not mentioned in a death certificate, thus making it impossible to link them with a birth certificate. However, when the partner of the deceased could be used as a second person in the linkage with the marriage certificate, matches between a death and marriage certificate were made. In combination with a link between a birth and a marriage certificate, the link between a birth and death certificate could be deduced (B links M, M links D, => B links D).

All these matching operations created multiple links (so-called overlinks). Certificates with multiple links were flagged and not linked, unless the composed Levenshtein values showed significant differences (e.g., 1 compared with 7 or 8). In that case a choice was made for the match with the lowest Levenshtein value and flagged as such.

6.3.3 SYNCHRONIZATION OF THE PERSON IDENTIFIERS

All persons from the civil certificates entered the dataset with their own identifiers which are always kept in the release as well. However, it may occur that the linkage information tells us for example that the person number 1203048 in the birth certificates is the same person as the person with number 42382209 in the marriage certificates. Then, we need to synchronize the identification numbers and create a kind of global identifier, Id, which is for each release. Synchronization of the identifying keys influences all generations. A child in a birth certificate may be a bride in a marriage certificate and a parent in the next generation of birth, death and marriage certificates. So, the link between the generations is made through the marriage certificates. But that also implies that the synchronization of the identifiers must start with the marriage certificates to make sure that the same parents in the birth or death certificates end up with the same identifiers. The matching has been done in the form of pedigrees, going backwards while we need life courses and families that start at the beginning of their life cycle. This implies, that the pedigrees need to be 'toppled' into family tree systems.

Synchronizing persons from the marriage certificates

To make the synchronization process feasible two extra steps are necessary: a) the pedigree system has to be transformed into a family tree system, and b) remarried children need to be identified.

The conversion from a pedigree system into a family tree system was done by way of the following steps:

- 1 Define the level of the family tree as generation "1" if there is no link with a parental marriage;
- 2 Define the family tree as generation "2" if there is a link with a previous marriage with generation level "1";
- 3 Repeat step 2 up to generation level 7 or more, which is the limit and occurs only four times in the Zeeland dataset.

This procedure looks more straightforward than it is. One needs to realize that although a bride or a groom has only one parental couple, they have two grandparental couples, four great-grandparental couples etc. This implies that in numbering the generations, different levels will apply to a person depending on the path backwards. For example, in one marriage the bride may have generation level 3 linking along the father line back to the grandparents and level 2 in case the mother line shows no further links backwards. This implies that at the second level we have four marriage lines to follow: the mother line (bride -> mother), the father line (groom -> father), the diverting mother line (bride -> father) and the diverting father line (groom -> mother). In the table MARRIAGE_LINES the levels of the first two lines are represented in the field Family_tree_level; the last two in the field Family_tree_level_A. For the third level we could have doubled this system again, but we abstained from this, not wanting to make the system too complicated.

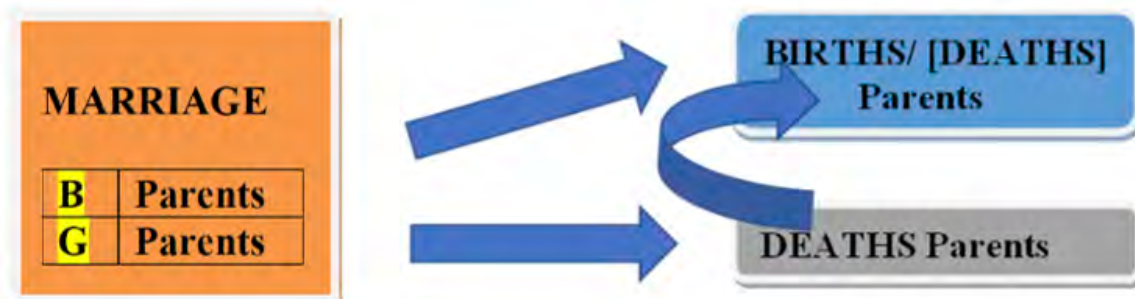
Another issue is that remarried children need to be identified. There are brides and grooms who marry more than once and initially have different identification numbers. If they are not identified as one and the same person they will be seen as siblings in the dataset. At generation level 2 and higher remarried persons are identified within the context of the parental marriage, in which they can be matched through their first name. In case of equal first names, the identification number is synchronized. At generation level 1 we created 'shadow marriages' (see Section 6.3.2, but now on the basis of marriage certificates) on which basis we could match on first names and synchronize the identifiers of remarried persons.

In a final step the identifiers of the bride and groom of generation 1 replace the identifiers of the parents of generation 2, etc. up until generation 5 replaces the identifiers of the parents of the 6th generation.

Synchronizing persons from the birth and death certificates

In first instance, the identifiers of the deceased, mother and father of the death certificate were replaced by those of the birth certificate by way of the linkage itself. All persons from the death certificates that were not linked kept their original identifiers. Eventual partners of the deceased always kept their own number because they were not included in the birth certificates.

Figure 10 *Synchronization scheme parents' birth and death certificates*



In a second step the identifiers of the bride and groom of the parental marriage certificate were used to replace the respective identifiers of the parents in the birth and death certificates (see Figure 10). On the basis of the linkage between the births/deaths and their own marriage certificates the identifiers of the births and the deaths in the table BIRTH_DEATH were replaced by the ones of the bride and the groom. In case of parents from generation level 1 (whose own marriage certificate has not been found), the identifiers of these parents were equalized with those from the birth or death certificates.

In a final step, on the basis of the links between the death and marriage certificates the partners in the death certificates were synchronized with the identifiers in the marriage certificate. In case of multiple marriages, the death certificates usually include the last partner. Since only those death certificates are used that have been linked with the marriage certificates including a link with the partner, possible false links are logically impossible.

6.3.4 RESULTS

Three regions were matched separately: the combination of the provinces of Groningen and Drenthe as well as Limburg and Zeeland. Table 10 presents the number of included certificates and the results of the matching and identifying process for each region. Almost 6 million certificates are included which stands for about 20 million person appearances. Matching was done along the lines of the releases as discussed in Sections 6.1. and 6.2 with the exception that the matching was limited to the certificates from the specific region.

All in all, we identified just under 8 million different persons in these three integrated systems. Some of these persons are combined into 407.435 families. With an average number of 4.29 children and 2 parents, which means that about 2.56 million unique persons are involved in a family structure defined as a married couple with at least one known child. Given the total of 7.99 million, it seems that 5.5 million persons are lacking. These are the persons that are included in the certificates that could not be linked. But this total of 5.5 million is seriously exaggerated because identical persons that are not linked within a family structure are not identified as such and are counted more than once.

The combination of the two provinces of Groningen and Drenthe has about 50% more indexed certificates than the other two provinces Zeeland and Limburg. In terms of linkage results between birth and death certificates there is no big difference between Groningen/Drenthe and Zeeland. Limburg has a much lower result, despite the relatively high number of included death certificates. The main reason for this result is the shape of the province having a much longer border with other provinces and Belgium and Germany than the other ones. This implies that we could expect that there was more in- and outmigration. And during the first half of the 20th century the coal area of Limburg attracted many persons from outside the province (Langeweg, 2012).

Table 10 *Integrated linking results for three areas: Groningen/Drenthe, Limburg and Zeeland*

	Groningen/ Drenthe	Zeeland	Limburg	Total
Number of included birth certificates	1,061,614	698,361	761,857	2,521,832
Number of included death certificates	1,043,926	650,728	843,413	2,538,067
Number of included marriage certificates	365,672	193,793	212,399	771,864
Number of linked birth/death	567,333	368,517	326,818	1,262,668
% of included births	53.4	52.8	42.9	50.1
% of included deaths	54.3	56.6	38.7	49.7
Number of linked brides and grooms with parental marriages	465,650	227,604	169,538	862,792
% of linked marriage lines	63.8	58.7	39.9	55.9
Number of linked births with parental marriage	835,081	511,647	402,933	1,749,661
% of included births	78.7	73.3	52.9	69.4
Number of linked deaths with parental marriage	585,120	351,058	305,468	1,241,646
% of included deaths	56.1	53.9	36.2	48.9
Number of unique persons	2,891,468	1,939,954	3,160,298	7,991,720
Number of families	201,882	106,082	99,471	407,435
Average number of children/family	4.14	4.82	4.05	4.29
Number of three generation pedigrees	246,855	109,847	83,293	439,995
Number of four or more generation pedigrees	153,555	53,442	34,170	241,167

Explanation: Number of death certificates Zeeland and Limburg include lifeless reported certificates (respectively $n=40,786$ and $n=52,068$). Lifeless reported cases are linked with marriage certificate of the parents (but are lacking a birth certificate). Families are defined as marriages with at least one identified child. Three-generation structures are pedigrees with at least two linked marriage certificates; a distinction has been made between a) the first three generations and b) 'doubling structures' in case of more than three generations (a sixth generation family structure contains four overlapping three-generation structures).

The results for links with the marriages show the same pattern of a relatively bad performance of Limburg. Especially for the pedigrees and links of the births with the parental marriages Groningen/Drenthe also shows a better result than Zeeland with a positive difference of about 5%. The average number of children per family is in line with what one would expect for the 19th century which was on average 4.7 children per family (van den Berg et al., 2021; also in line with Dribe et al. (2017) and Engelen (2009; p. 174) who came to the same result on the basis of the census outcomes).

6.3.5 SOME REMARKS ABOUT THE RESULTS

Matching certificates to reconstruct life courses, pedigrees, family trees, families, etc. from a limited area and time period, implies several 'data leaks' and inconsistencies, mainly because of the following reasons:

- 1 Persons could have emigrated to another area;
- 2 Persons could immigrate from another area;
- 3 Not all certificates are matched because of insufficient identifying information;
- 4 Certificates are matched in an ambiguous way because the identifying information is not accurate enough;
- 5 Persons cannot be matched because the certificate to be matched does not exist (before 1812) or is not indexed yet;
- 6 Bugs in the matching software;
- 7 Inconsistencies in one generation may have consequences for the family trees that have been constructed.

In the releases, several fields are included that describe the way the data have been matched. These flags may be used to make selections from the dataset to test how robust the outcomes of the statistical analyses are. Ultimately, it is the researcher who is responsible for the way the data are used.

Since the marriage certificates are linked with both birth and death certificates, it is possible to check on triangle problems. It turns out that many death certificates are linked with marriage certificates and not with birth certificates, where these birth certificates are linked with the same marriage certificate. In the case of death certificates of persons born before 1850, in which names of parents are often of poor quality if mentioned at all, they could easily be linked to their own marriage certificates on the basis of the names of partners.

Reshaping the pedigrees into family trees, is a kind of toppling of the pedigree system. It is an essential step because most of the generational analysis should be done from the perspective of the beginning of a family line (not of the end), especially when the system is to be extended with other certificates (e.g., of the children of the couples). There is also a practical problem: because each line in the pedigree will result in a different generation level for the last generation, one cannot simply fix the generation level for one marriage line in a marriage certificate. The more generations are involved in such a system the more complicated this will become.

6.3.6 DATASETS FOR ANALYSIS

The *integrated* datasets of births, marriages and deaths with created families and multigenerational links are not sufficient to be immediately usable for research. For the Zeeland release two types of datasets suitable for statistical analysis were created. The first one was a rectangular-type structure that was constructed within the context of the project *Genes, Germs and Resources* (<https://www.nwo.nl/en/projects/360-53-180-0>; Mourits et al., 2020). The second was a conversion of the format of the Zeeland release into the format of the Intermediate Data Structure (Alter & Mandemakers, 2014).

The database which was constructed for the project *Genes, Germs and Resources* (LINKS-gen; see Mourits et al., 2020), served several goals. The first one was to create more explicit family links than were provided in the Zeeland release and to improve and extend dates of birth, last observation and other variables. The second one was to reformat the design into a so-called pedigree format.

Through better integration of unlinked newborn and deceased persons that were linked to the same parents, a more consistent dataset could be created. Also, several data improvements were applied. For example, the conversion of ages at a specific moment into birth ranges, fields were created for up to five marriages and newborns who were reported dead on registration lacked a date of birth which was included as the date of death. Other improvements made twins explicit, added dates of last observation and flagged complete cases. To retain all data and relational information of a person on one record the database was restructured into a so-called pedigree structure. This format structures the data in such a way that each record includes the identification number of a person, the identification numbers of his or her parents (if known), the sex and all other variables. Families and familial relationships are defined through the father and mother. Through restructuring the dataset, hidden links between persons were also made explicit. By this operation a new structure was created, making it also much easier for researchers to select their case for a specific analysis (Mourits et al., 2020).

The conversion of the Zeeland release into the IDS-format was relatively easy, mainly because the number of variables, or types in IDS-grammar, is quite limited (Mandemakers & Laan, 2017, IDS version). In the INDIVIDUAL table we have, sex, occupations, and the date and location of birth, marriage and death. Relations that are established in the INDIV_INDIV table are those between children and parents (including in-law relationships) and marriage couples. The nature and location of the certificates were used as the lowest level in the contextual system. An alternative could have been the use of the "Union" concept as lowest level as Klancher Merchant and Alter (2017) have done, but this approach was not necessary given the nature of the research for which the IDS dataset was developed. It concerned research into intergenerational effects of infant mortality in which four other databases were involved. All were structured into the IDS and reshaped in datasets ready for statistical analysis by a common Stata script (Quaranta, 2018; van Dijk & Mandemakers, 2018).

7 SUMMARY AND CONCLUSIONS

In this paper we explained the construction of the LINKS database using the indices of the civil certificates as collected by the Dutch Family Center and published on the website *WieWasWie*. Presently, over 40 million certificates and 120 million appearances of persons have been included in this index by hundreds of volunteers working over the last twenty years to create an electronic index of the civil certificates as soon as they become public.

Two matching systems have been developed within the HSNDB environment. The first one is a query system based on SQL queries selecting the data from the MySQL-database in which the matching queries are only part of a wider environment directed at the standardizing, cleaning, enriching and outputting of the LINKS data. Using the Zeeland marriage certificates as an example, we showed the excellent quality of the data material in general. The difference between the result of the exact matching and the least restricted, yet acceptable alternative in matching was only 38,338 matches (23.5%). Moreover, 80% of all matches were exact matches, thanks to the legal structure of civil registration that standardized the names and a legal and administrative structure that kept the birth name of the females alive with and after marriage.

However, on large datasets these matching queries are slow and the end user has no direct influence on the matching alternatives, which are set by the database manager unless special requests are made. For these reasons a second matching system, *burgerLinker*, was developed, based on knowledge graphs which can be run independently of the LINKS environment.

In both cases, matching is a first step for the creation of a linked dataset that can be used for research. We explained the construction of several relatively recent data releases. On a national scale, we released a pedigree system based on all marriage certificates and a dataset in which the births are linked with the marriages of their parents, forming families. On a regional level we created three separate releases for the provinces of Zeeland, Limburg and Groningen/Drenthe. Here, we combined birth, death and marriage certificates to create three-generation families. One of the issues for which we found a solution was the identification of unique persons in this three-generation system.

The LINKS project started in 2010. Since then, over 40 releases have been produced resulting in over 50 publications including several dissertations (Mandemakers & Kok, 2020). We expect that in the future, more releases will be made by HSNDB or by individual users of *burgerLinker*. Increasingly, researchers are using *burgerLinker* to link large collections of individual-level data. Military, inheritance tax and income tax registers have all proven to be important sources for future research. LINKS continues on the path set out by previous generations of historical demographers, creating new options for generations to come.

REFERENCES

- Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies*, 1, 1–26. doi: [10.51964/hlcs9290](https://doi.org/10.51964/hlcs9290)
- Bloothoof, G., van Boheemen, J., & Schraagen, M. (2016). Historical life cycle reconstruction by indexing. *Workshop Data Linkage: Techniques, Challenges and Applications at Isaac Newton Institute for Mathematics*. Cambridge UK. Retrieved from https://www.gerritbloothoof.nl/Publications/Cambridge_Bloothoof_etal.pdf
- Bloothoof, G., & Onland, D. (2016). Multiple first names in the Netherlands (1760–2014). *Names*, 64(1), 3–18. doi: [10.1080/00277738.2016.1118860](https://doi.org/10.1080/00277738.2016.1118860)
- Bloothoof, G., Onland, D., Reynaert, M., Depuydt, K., Schoonheim, T., Fannee, M., & Noordzij, J. (2020). *NAMES Corpus*. Retrieved from <https://taalmaterialen.ivdnt.org/download/names-corpus/>
- Bloothoof, G., & Schraagen, M. (2015). Learning name variants from inexact high-confidence matches. In: G. Bloothoof, P. Christen, K. Mandemakers, & M. Schraagen (Eds.). *Population Reconstruction* (pp. 87–110). Cham: Springer. doi: [10.1007/978-3-319-19884-2_4](https://doi.org/10.1007/978-3-319-19884-2_4)
- Bourdieu, J., Kesztenbaum, L., Postel-Vinay, G., & Tovey, J. (2014). The TRA project, a historical matrix. *Population (English Edition)*, 69(2), 191–220. doi: [10.3917/popu.1402.0217](https://doi.org/10.3917/popu.1402.0217)

- Christen, P., Vatsalan, D., & Fu, Z. (2015). Advanced record linkage methods and privacy aspects for population reconstruction — A survey and case studies. In: G. Bloothoof, P. Christen, K. Mandemakers, & M. Schraagen (Eds.). *Population Reconstruction* (pp. 87–110). Cham: Springer. doi: [10.1007/978-3-319-19884-2_5](https://doi.org/10.1007/978-3-319-19884-2_5)
- Dillon, L., Amorevieta-Gentil, M., Caron, M., Lewis, C., Guay-Giroux, A., Desjardins, B., & Gagnon, A. (2018). The *Programme de recherche en démographie historique*: Past, present and future developments in family reconstitution. *The History of the Family*, 23(1), 20–53. doi: [10.1080/1081602X.2016.1222501](https://doi.org/10.1080/1081602X.2016.1222501)
- Dribe, M., Breschi, M., Gagnon, A., Gauvreau, D., Hanson, H. A., Maloney, Th. N., Mazzone, S., Molitoris, J., Pozzi, L., Smith, K. R., & Vézina, H. (2017). Socio-economic status and fertility decline: Insights from historical transitions in Europe and North America. *Population Studies*, 71(1), 3–21. doi: [10.1080/00324728.2016.1253857](https://doi.org/10.1080/00324728.2016.1253857)
- Dupâquier, J., Kessler, D. (Eds.). (1992). *La société française au XIXe siècle. Tradition, transition, transformations* [French society in the XIX century. Tradition, transition, transformation]. Paris: Fayard.
- Engelen, Th. (2009). *Van 2 miljoen naar 16 miljoen mensen. Demografie van Nederland, 1800–nu* [From 2 million to 16 million people. Demography of the Netherlands, 1800–present]. Amsterdam: Boom. Retrieved from <http://hdl.handle.net/2066/78820>
- Gerritzen, D. (1998). Voornamen in Zeeland [Firstnames in Sealand]. In: K. Mandemakers, O. Hoogerhuis, & A. de Klerk (Eds.). *Over Zeeuwse mensen. Demografische en sociale ontwikkelingen in Zeeland in de negentiende en twintigste eeuw* [Special issue]. *Zeeland*, 7(3), 104–115.
- Goeken, R., Huynh, L., Lynch, T.A., & Vick, R. (2011). New methods of census record linking. *Historical Methods*, 44(1), 7–14. doi: [10.1080/01615440.2010.517152](https://doi.org/10.1080/01615440.2010.517152)
- Henry, L., & Fleury, M. (1956). *Manuel de dépouillement et d'exploitation de l'état civil ancien* [Manual to analyse and exploit the ancient civil registration]. Paris: INED.
- Huijsmans, D. P. (2020). *HSN Gazetteer* [Data set]. Retrieved from <https://hdl.handle.net/10622/ZDT2DJ>
- Klancher Merchant, E., & Alter, G. (2017). IDS Transposer: A users guide. *Historical Life Course Studies*, 4, 59–96. doi: [10.51964/hlcs9339](https://doi.org/10.51964/hlcs9339)
- Kok, J. (1991). *Langs verboden wegen. De achtergronden van buitenechtelijke geboorten in Noord-Holland 1812–1914* [Along forbidden roads. Background of illegitimate births in North-Holland 1812–1914]. Hilversum: Verloren.
- Lambert, P. S., Zijdeman, R. L., van Leeuwen, M. H. D., Maas, I., & Prandy, K. (2013). The construction of HISCAM: A stratification scale based on social interactions for historical comparative research. *Historical Methods*, 46(2), 77–89. doi: [10.1080/01615440.2012.715569](https://doi.org/10.1080/01615440.2012.715569)
- Langeweg, S. (2012). Werving, herkomst en binding van mijnwerkers [Recruitment, origin and binding of miners]. In: Knotter, A. (Ed.), *Mijnwerkers in Limburg. Een sociale geschiedenis* (pp. 100–138). Nijmegen: Vantilt.
- Mandemakers, K. (2000). The Netherlands. Historical Sample of the Netherlands. In: P. Kelly Hall, R. McCaa, & G. Thorvaldsen (Eds.), *Handbook of international historical microdata for population research* (pp. 149–178). Minneapolis: Minnesota Population Center. Retrieved from https://international.ipums.org/international/resources/microdata_handbook/1_10_netherlands_ch11.pdf
- Mandemakers, K. (2023, January 20). “You really got me”. *Ontwikkeling en toekomst van historische databestanden met microdata* [Development and future of historical databases with microdata] (Valedictory speech). Erasmus University, Rotterdam, the Netherlands. doi: [10.25397/eur.23256467](https://doi.org/10.25397/eur.23256467)
- Mandemakers, K., & Dillon, L. (2004). Best practices with large databases on historical populations. *Historical Methods*, 37(1), 34–38. doi: [10.3200/HMTS.37.1.34-38](https://doi.org/10.3200/HMTS.37.1.34-38)
- Mandemakers, K., Hornix, J., Mourits, R. J., Muurling, S., Boter, C., van Dijk, I. K., Maas, I., Van de Putte, B., Zijdeman, R. L., Lambert, P., van Leeuwen, M. H. D., van Poppel, F., & Miles, A. (2020). *HSN standardized, HISCO-coded and classified occupational titles, HSN release 2020.02* [Data set]. Retrieved from <https://hdl.handle.net/10622/88ZXD8>
- Mandemakers, K., & Kok, J. (2020). Dutch lives. The Historical Sample of the Netherlands (1987–): Development and research. *Historical Life Course Studies*, 9, 69–113. doi: [10.51964/hlcs9298](https://doi.org/10.51964/hlcs9298)
- Mandemakers, K., & Laan, F. (2017). *LINKS Zeeland linked dataset (Marriages, births and deaths), province of Zeeland, Release 2017_02, including IDS format* [Data set].
- Mandemakers, K., & Laan, F. (2018). *LINKS Groningen-Drenthe linked dataset (Marriages, births and deaths), Release 2018_01* [Data set].

- Mandemakers, K., & Laan, F. (2019). *LINKS dataset WieWasWie Limburg, linked civil certificates (Births, deaths and marriages), Release 2019.02* [Data set].
- Mandemakers, K., & Laan, F. (2020a). *LINKS dataset linked births and marriage certificates parents, the Netherlands, Release 2020.01* [Data set].
- Mandemakers, K., & Laan, F. (2020b). *LINKS dataset linked marriages, the Netherlands, 1796–1943, Release 2020.03 (n=4,158,388), Also a version including first names of bride/groom and parents, Release 2020.03_f* [Data set].
- Mourits, R. J., Boonstra, O., Knippenberg, H., Hofstee, E. W., & Zijdemans, R. L. (2016). *Historische Database Nederlandse Gemeenten* [Data set]. Retrieved from <https://hdl.handle.net/10622/RPBVK4>
- Mourits, R. J., van Dijk, I. K., & Mandemakers, K. (2020). From matched certificates to related persons. *Historical Life Course Studies*, 9, 49–68. doi: [10.51964/hlcs9310](https://doi.org/10.51964/hlcs9310)
- Nault, F., & Desjardins, B. (1989). Computers and historical demography: The reconstitution of the early Québec population. In: P. Denley, S. Fogelvik, & Ch. Harvey. *History and computing, II*. (pp. 143–148). Manchester: Manchester University Press.
- Quaranta, L. (2018). Program for studying intergenerational transmissions in infant mortality using the Intermediate Data Structure (IDS). *Historical Life Course Studies*, 7, 11–27. doi: [10.51964/hlcs9287](https://doi.org/10.51964/hlcs9287)
- Oosten, M. (2008). *Verleden namen. Familieverbanden uit Genlias-data* [Names from the past. Family structures from GENLIAS data] (Unpublished master's thesis). LIACS and IISG, Leiden.
- Raad, J., Mourits, R. J., Rijpma, A., Schalk, R., Zijdemans, R. L., Mandemakers, K., & Meroño-Peñuela, A. (2020). Linking Dutch civil certificates. *3rd Workshop on Humanities in the Semantic Web (WHiSe) conference proceedings*. Heraklion, Greece. Retrieved from <https://ceur-ws.org/Vol-2695/paper6.pdf>
- Schraagen, M. (2014). *Aspects of record linkage* (PhD thesis). Leiden University. Retrieved from <http://hdl.handle.net/1887/29716>
- Schulz, K. U., & Mihov, S. (2002). Fast string correction with Levenshtein automata. *International Journal on Document Analysis and Recognition*, 5(1), 67–85. doi: [10.1007/s10032-002-0082-8](https://doi.org/10.1007/s10032-002-0082-8)
- Séguy, I. (2001). *La population de la France de 1670 à 1829: l'Enquête Louis Henry et ses données* [The population of France from 1670 to 1829: The Louis Henry survey and its data]. Paris: INED.
- Séguy, I. (2016). The French school of historical demography (1950–2000). In: A. Fauve-Chamoux, I. Bolovan, & S. Sogner (Eds.). *A global history of historical demography. Half a century of interdisciplinarity* (pp. 257–276). Bern: Peter Lang.
- Song, X., & Campbell, C. D. (2017). Genealogical microdata and their significance for social science. *Annual Review of Sociology*, 43, 75–99. doi: [10.1146%2Fannurev-soc-073014-112157](https://doi.org/10.1146%2Fannurev-soc-073014-112157)
- Stead, W. W., Hammond, W. E., & Straube, M. J. (1982, November). A chartless record — Is it adequate? *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 89–94. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2580254/>
- van Boheemen, J. (2016). *Assembling the pages. A sorting-based approach to historical record linkage*. (Bachelor's thesis). Universiteit Utrecht. Retrieved from <https://studenttheses.uu.nl/handle/20.500.12932/23905>
- van den Berg, N., van Dijk, I. K., Mourits, R. J., Slagboom, P. E., Janssens, A. A. P. O., & Mandemakers, K. (2021). Families in comparison: An individual-level comparison of life-course and family reconstructions between population and vital event registers. *Population Studies*, 75(1), 91–110. doi: [10.1080/00324728.2020.1718186](https://doi.org/10.1080/00324728.2020.1718186)
- van der Bie, R. J. (1995). "Een doorlopende grote roes". *De economische ontwikkeling van Nederland 1913–1921* ["A continuous big Rush". The economic development of the Netherlands, 1913–1921]. Amsterdam: Tinbergen Institute research Series.
- van der Bie, R. J., & Smits, J. P. (Eds.). (2000). *Tweehonderd jaar statistiek in tijdreeksen, 1800–1999* [Two hundred year statistics in time series, 1800–1999]. Amsterdam: Stichting Beheer IISG.
- van Dijk, I. K., & Mandemakers, K. (2018). Like mother, like daughter. Intergenerational transmission of infant mortality clustering in Zeeland, the Netherlands, 1833–1912. *Historical Life Course Studies*, 7, 28–46. doi: [10.51964/hlcs9286](https://doi.org/10.51964/hlcs9286)
- van Galen, C. W. (2019). Creating an audience: Experiences from the Surinamese slave registers crowdsourcing project. *Historical Methods*, 52(3), 178–194. doi: [10.1080/01615440.2019.1590268](https://doi.org/10.1080/01615440.2019.1590268)
- van Galen, C. W., Mourits, R. J., Rosenbaum-Feldbrügge, M., A.B., M., Janssen, J., Quanjer, B., van Oort, Th., & Kok, J. (forthcoming). Slavery in Suriname: A reconstruction of life courses, 1830–1863. *Historical Life Course Studies*.

- van Leeuwen, M. H. D., & Maas, I. (2011). *HISCLASS. A Historical International Social Class Scheme*. Leuven: Leuven University Press.
- van Leeuwen, M. H. D., Maas, I., & Miles, A. (2002). *HISCO. Historical International Standard Classification of Occupations*. Leuven: Leuven University Press.
- van Zanden, J. L., & Griffiths, R. T. (1989). *Economische geschiedenis van Nederland in de 20e eeuw* [Economic history of the Netherlands in the 20th century]. Utrecht: Uitgeverij Het Spectrum.
- Vézina, H., & Bournival, J.-S. (2020). An overview of the BALSAC population database. Past developments, current state and future prospects. *Historical Life Course Studies*, 9, 114–129. doi: [10.51964/hlcs9299](https://doi.org/10.51964/hlcs9299)
- Vulsma, R. F. (1988). *Burgerlijke stand en bevolkingsregister* [Civil registration and population register]. 's-Gravenhage: Centraal Bureau voor de Genealogie.
- 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. doi: [10.1017/CBO9780511660344](https://doi.org/10.1017/CBO9780511660344)

APPENDIX A OVERVIEW OF ERROR REPORTS

This appendix is an overview of all types of error reporting (logic, completeness and errors). The field Type provides a reference to the "error message". The field Class consists of three values: "FT" for Fout ('error'), which means that some error has taken place; "WA" for Waarschuwing ('warning') which means that a value could be wrong but there is a chance that it is not a problem, which is typically for values not yet standardized; "NB" stands for a value that needs to be checked, but which is not necessarily a mistake (depends on the archive). The field Content provides the error message that is reported for the error type, e.g., type 41 returns a message like "Non authorized occupation: shoematter", "No standard; standard_code= x". The messages are delivered with information identifying the original source.

Type	Class	Content
1	FT	Double entry of the original registration
2	FT	Registration without a registration date
3	FT	Registration without defining one or more roles
4	FT	One of more than two entries with same Registration Details (Type, Location, Year and Sequence)
21	WA	Non authorized religion: No standard; standard_code= "x"
23	WA	Invalid religion: No standard; standard_code= "n"
25	WA	Invalid religion: Standard present; standard_code= "u"
29	FT	Standard_code not valid
31	WA	Non authorized gender: No standard; standard_code= "x"
33	WA	Invalid gender: No standard; standard_code= "n"
35	WA	Invalid gender: Standard present; standard_code= "u"
39	FT	Standard_code not valid
41	WA	Non authorized occupation: No standard; standard_code= "x"
43	WA	Invalid occupation: No standard; standard_code= "n"
45	WA	Invalid occupation: Standard present; standard_code= "u"
49	FT	Standard_code not valid
51	WA	Non authorized registratietype: No standard; standard_code= "x"
53	WA	Invalid registration type: No standard; standard_code= "n"
55	WA	Invalid registration type: Standard present; standard_code= "u"
59	FT	Standard_code not valid
61	WA	Non authorized gender or status: No standard; standard_code= "x"
63	WA	Invalid gender or status: No standard; standard_code= "n"
65	WA	Invalid gender or status: Standard present; standard_code= "u"
68	FT	Civil Status: suggests a gender wich is inconsistent with the gender of this person
69	FT	Standard_code not valid
71	WA	Non authorized suffix: No standard; standard_code= "x"
73	WA	Invalid suffix: No standard; standard_code= "n"
75	WA	Invalid suffix: Standard present; standard_code= "u"
79	WA	Standard_code not valid
81	WA	Non authorized title or prefix: No standard; standard_code= "x"
83	WA	Invalid title or prefix: No standard; standard_code= "n"
85	WA	Invalid title or prefix: standaard aanwezig; standard_code= "u"
89	FT	Standard_code not valid
91	WA	Non authorized location: No standard; standard_code= "x"
93	WA	Invalid location: No standard; standard_code= "n"

Type	Class	Content
95	WA	Invalid location: Standard present; standard_code= "u"
99	FT	Standard_code not valid
102	NB	Restant opmerking:
103	NB	Not valid combination of role: [rol] and date: [date]
104	NB	Invalid function code: from table ref_date_minmax
105	FT	Could not find all info in reference table "ref_date_minmax" to calculate minmax: <>
106	FT	Function minMax/MainAge cannot find record in ref_date_minmax
107	FT	Duplicate role within one registration
111	NB	Sequence number not present
112	NB	Sequence number is not numeric:
113	FT	Sequence number occurred twice:
114	FT	Missing Sequence number (previous number is lacking):
115	FT	There are more than 100 records for a specific source per year per municipality but december is missing; 100:12 rule
141	WA	Non authorized role: No standard; standard_code= "x"
142	FT	Invalid role: In combination with registration type
143	WA	Invalid role: No standard; standard_code= "n"
145	WA	Invalid role: Standard present; standard_code= "u"
149	FT	Standard_code not valid
201	FT	Constructed (from events) registration date is invalid:
202	WA	Date of registration based only on the year of the registration
203	WA	Invalid registration_date, but reconstructable
204	FT	Components registration date are are invalid
205	FT	No Registration date and registration_date is not constructable
206	WA	Registration date and registration elements unequal
211	FT	Invalid Birth date:
221	FT	Invalid Marriage date:
231	FT	Invalid Death date:
241	FT	Age in days is out of range (0-99):
242	FT	Age in weeks is out of range (0-49):
243	FT	Age in months is out of range (0-49):
244	FT	Age in years is out of range (0-114):
251	WA	Non authorized literal age: No standard; standard_code= "x"
253	WA	Invalid literal age: No standard; standard_code= "n"
255	WA	Invalid literal_age: Standard present; standard_code= "u"
259	FT	Standard_code not valid
261	WA	Content Age_literal: conflicts with Age_year:
262	WA	Content Age_literal: conflicts with Age_month:
263	WA	Content Age_literal: conflicts with Age_week:
264	WA	Content Age_literal: conflicts with Age_day
265	FT	Content Age_year: where role is parent or partner
266	FT	Minimum Age: larger than Maximum Age:
267	FT	Content Age_literal: conflicts with Role is "Kind"
271	FT	Missing newborn in birth registration

Type	Class	Content
272	FT	Missing bride in marriage registration
273	FT	Missing groom in marriage registration
274	FT	Missing deceased in death registration
281	WA	More than one newborn in birth registration
282	WA	More than one bride in marriage registration
283	WA	More than one groom in marriage registration
284	WA	More than one deceased in birth registration
1000	FT	Invalid familie name: standard present; standard_code= "u"
1001	FT	Person has no family name
1002	WA	Family name: uncleaned familyname does not exists in ref_file
1003	FT	Family name: contains two or more serried spaces (automatically corrected)
1004	FT	Family name: contains invalid character (automatically corrected)
1005	FT	Invalid familie name: No standard; standard_code= "n"
1006	FT	Invalid family name: contains suffix
1007	FT	Invalid family name: contains an alias
1008	FT	Invalid family name: contains prefix/title
1009	WA	Non authorized family name: No standard; standard_code= "x"
1010	FT	Standard_code not valid
1011	WA	Famillyname includes string without spaces
1012	FT	Famillyname includes prefix as a suffix
1100	FT	Invalid first name: Standard present; standard_code= "u"
1101	FT	Person has no first name
1104	FT	First name: contains invalid character (automatically corrected)
1105	FT	Invalid first name: No standard; standard_code= "n"
1106	FT	Invalid first name: contains suffix
1107	FT	Invalid first name: contains alias
1108	FT	Invalid first name: contains prefix/title
1109	WA	Non authorized first name: No standard; standard_code= "x"
1110	FT	Standard_code not valid
1111	WA	Firstname includes string without spaces
1112	WA	Firstname includes embedded capital:
1113	WA	Firstname includes embedded slash:
1114	WA	Firstname includes embedded HTML break:
1203	WA	Prefix, postfix or alias: contains two or more serried spaces (automatically corrected)
1204	WA	Prefix, postfix or alias: contains invalid character (automatically corrected)
1211	WA	Prefix, postfix or alias: includes string without spaces

APPENDIX B THE CIVIC REGISTRY MODEL

In this appendix the classes and schemas of the Civic Registry Model (CIV) showed in Figure 4 are described in a more formal way.

The CIV-model is composed of three parts:

1. Person (blue)

This part is only composed of the class schema:Person, representing the individuals described in the civil registries. An instance of this class must have a unique identifier (civ:personID), a first name (schema:givenName), and a last name (schema:familyName). All these properties are required for linking persons. In addition, for improving the accuracy and the speed of linking, adding the gender (schema:gender) of every individual is recommended.

2. Events (green)

We make a distinction between three different types of events: civ:Birth, civ:Marriage, and civ:Death. These three types of events are all sub-types of the general class civ:Event. Being sub-type of civ:Event means that these three classes inherit the properties of their general class, i.e., each instance of the class civ:Birth, civ:Marriage, and civ:Death can have the five relations that are associated with civ:Event. Out of these five relations, only two are required for linking: a unique event/registration identifier (civ:registrationID) and the date of an event (civ:eventDate). The remaining three optional relations are used for indicating the date of registration (civ:registrationDate), its location (civ:registrationLocation) and the event location (civ:eventLocation). In this model, a distinction is made between the date/location of an event and the date/location of its registration in the civil registries, as certain civil registrations can be produced in different dates and locations from where the life event happened.

In addition, each of these three types of event has different relations associated to it:

civ:Birth

An instance of this class can have the three properties: civ:newborn, civ:mother, and civ:father. For linking, all information regarding the newborn must be present in a birth event, in addition to at least one of their parents.

civ:Marriage

An instance of this class can also have the six properties: civ:bride, civ:motherBride, civ:fatherBride, civ:groom, civ:motherGroom, civ:fatherGroom. For linking, all information regarding the bride and groom must be present in a marriage event, in addition to at least one parent for each of the bride and groom.

civ:Death

An instance of this class can also have the four properties: civ:deceased, civ:partner, civ:mother civ:father. For linking, all information regarding the deceased must be present in a death event, in addition to at least one of their parents.

3. Location (yellow)

The final part describes the location where each life event has happened and the location where it was registered. In this part, information regarding the municipality, the province, the region, and the country can be available. This part is completely optional, as none of the information regarding the locations of the events and their registrations are used for linking.

Historical Population Database of Transylvania

Sources, Particularities, Challenges, and Early Findings

Luminița Dumănescu	Centre for Population Studies, Babeş-Bolyai University, Cluj-Napoca
Mihaela Hărăgus,	Centre for Population Studies, Babeş-Bolyai University, Cluj-Napoca
Angela Lumezeanu	Centre for Population Studies, Babeş-Bolyai University, Cluj-Napoca
Elena Crinela Holom	Centre for Population Studies, Babeş-Bolyai University, Cluj-Napoca
Nicoleta Hegedűs	Centre for Population Studies, Babeş-Bolyai University, Cluj-Napoca
Daniela Mârza	Center for Transylvanian Studies, Romanian Academy, Cluj-Napoca
Diana Covaci	Centre for Population Studies, Babeş-Bolyai University, Cluj-Napoca
Ioan Bolovan	Centre for Population Studies, Babeş-Bolyai University, Cluj-Napoca

ABSTRACT

The Historical Population Database of Transylvania (HPDT) is a research tool for population studies developed since 2014 at the Centre for Population Studies in Cluj-Napoca, financed by an SEE-Norway Grant. HPDT employs a source-oriented approach for recording data from the parish registers kept by the Transylvanian churches, focusing primarily on the main vital events such as births, marriages, and deaths. The data entry process was followed by the standardization of various information, such as names, occupations, locations and causes of death, thus allowing the initiation of a linkage process. The database has already been employed in a wide-ranging series of analyses conducted on datasets extracted from HPDT, which include infant and adult mortality, nuptiality and age at first marriage, social mobility, and the medicalization of childbirth. The wealth of information it includes will enable many more scientific investigations.

Keywords: Historical database, HPDT, Transylvania, Parish registers, Historical demography

DOI article: <https://doi.org/10.51964/hlcs12038>

© 2022, Dumănescu, Hărăgus, Lumezeanu, Holom, Hegedűs, Mârza, Covaci, Bolovan

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

Historical demography began its journey in Romania in the interwar period and grew in the shadow of other sciences. Since the 1960s, case studies and microlevel analyses based on the Henry and Fleury method of family reconstitution have enriched the Romanian historical demography. Only recently this research has developed a population database constructed according to the most recent methodological principles (Mandemakers & Dillon, 2004) and aiming for full compatibility with the latest version of the *Intermediate Data Structure* (IDS; Alter & Mandemakers, 2014). The Historical Population Database of Transylvania (HPDT) was developed by the Centre for Population Studies in Cluj-Napoca. Data entry started in 2014, after the project received financial support and was developed together with the Norwegian Historical Data Centre (Tromsø, Norway). A team of almost 20 people — researchers and data entry operators — started to collect, transcribe, standardize and link data from a sample of around 7% of the historical population of Transylvania, and they delivered a working database compatible with the existing international standards. After the project's completion in April 2017, researchers of the Centre for Population Studies continued to clean, standardize, link and add data to HPDT. The purpose of this paper is to present the Transylvanian database, one of the newest databases in Europe, focusing on its particularities and on the challenges encountered during its development and implementation.

Development of the database was based on the source-oriented approach and attempted to faithfully replicate all types of data found in the parish registers, including the smallest possible details. The database was constantly adapted to integrate new fields as soon as new types of data emerged, thus reflecting the complexity and the diversity of the sources.

The Historical Population Database of Transylvania (HPDT) provides a tool for researching demographic phenomena at the micro level in Transylvania, the northern part of present-day Romania (see Figure 5). The intimate details revealed by the newly available data allowed scholars to see a world that functioned differently than previously thought, when only macro-level data were available (census, aggregates). These sources shaped a social history which sheds new light on the old world — often considered obsolete and traditional — of a province situated at the periphery of the Austrian-Hungarian Empire (until it became part of Romania in 1919).

The main sources of the HPDT are the parish registers kept by churches and covering the period between 1850 and 1914–1920. These sources recorded the most important events in an individual life course from a demographic perspective. The architecture of the database was conceptualized by starting with the three principal registers: those for the baptism, marriage, and burial events. Betrothals were added to the database where they were available.

The structure of the database is based on the main events from the parish registers: births, marriages, burials. Each main table incorporates all the fields that can be found in the register for the respective event regardless of the denomination of the register. Each data entry form reflects the structure of the corresponding table from the database. Because of the complexity of the sources the architecture of the database steadily evolved during the project implementation.

The Historical Population Database of Transylvania is a relational database, implemented in MySQL, an open source database management system. There are two main components of HPDT: a research database on the one hand, and a public open access one on the other. The research component is built on three elements: a source database, a standard database and a third database which contains linked individuals. The public database (<http://hpdt.ro:4080>) is an open access website that offers insights on the Transylvanian population to a broader public. The structure of the database has been extensively described in the doctoral thesis of Angela-Cristina Lumezeanu (2019).

The following sections address the particularities of parish registers in Transylvania, sampling strategy and sample composition, principles of data entry, and ongoing development of the database structure in response to the heterogeneity of the information found in the parish registers. We follow the discussion of the database with an overview of the main results of research based on HPDT data.

2 SOURCES AND DATA ENTRY

2.1 SOURCES

The main sources of the HPDT are the parish registers that recorded life events such as birth of a child (in Baptism registers), wedding (in Marriages registers), engagements (in Betrothal registers) and deaths (in Burial registers). Until 1895, when Hungarian Law enforced the civil registration of the life events, these parish registers could be regarded as official records, and, despite all their limits and fragmentary character, they appear to be reliable sources for historical demography. The HPDT is also based on parish registers after 1895, since inclusion of civil certificates would have required a new structure of the database.

Church registers in Transylvania were written in several languages (Romanian, Hungarian, Latin, German, etc.), with different alphabets (Latin, Cyrillic, etc.), and came from diverse denominations: Orthodox, Greek-Catholic, Roman-Catholic, Reformed, Lutheran and Jewish. This diversity translated into different spellings of names and variation in the registers' headings, which required multiple adaptations of the database structure (discussed in next sections). The following Figures 1–4 give some examples of these sources, illustrating the structural heterogeneity of the parish registers.

Information about births was derived from the baptismal registers (see Figures 1 and 2). The information was organized in three sections with fields for data regarding the child, his parents and the baptism respectively. Some of these are text fields, for the name of the child and of the parents, the birth place, the baptism place, the parents' occupations and residence. In some parish registers the data was minimal with only the more usual categories of information including the date of the baptism, the name of the child, the parents' names (in some cases only the father is mentioned), the godparents, and the name of the priest who performed the baptism. Other sources held a wealth of additional information about the occupation of the parents, their age, their other relatives, the address of residence, the midwife, the age, status and occupation of the godparents, about vaccination (the date thereof, the name of the physician performing the inoculation) and even about the death of the baptized (sometimes filled in decades after the baptism record). In addition to the general categories of information, there were specific ones such as: legitimacy of the child, whether the child was the result of a single or multiple birth, if he or she was anointed or not (one of the sacraments).

Figure 1 Excerpt from a baptism register, 1863 (Latin)



Note: The columns contain the following information: the year, month, day for the birth and baptism, the first name of the baptised child, place of baptism, the parents of the child, their occupation and place of living, child's denomination, the name of the godparents, the name of the priest, vaccination date and reflections.

Figure 2 Excerpt from a baptism register, 1879 (Hungarian)

KERESZTELESI						ANYAKÖNYV.					
Folyt. szám	Ev és napja a keresztelésnek	MEGERESZTELTEK		Terveztettség	Dicsőítő-letölés	Szülők neve, vallás, állapota.	Lakhat. a ház számával	Keresztatyák a anyák neve, állapota.	Keresztelő-lelkész neve	Bálya neve	Eszrevételek
		neve	születés helye								
25	1879 Augusztus 25	25	30			Baetha Dániel/Anna/ Szász Bábel/ of földes	299	Szék György/ Horváth Susanna/ of földes	Kovács Sándor	Borbála Bábel	
26	1879 Szeptember 9	9	11			Károly/Anna/ János/ of földes	120	Házi Sándor/ Szabó Bábel/ of földes	Kovács Sándor	Anna	Édes anyja 1879. 9. 11. napján született 200/90
27	1879 Szeptember 8	8	14			János/Anna/ Szék György/ of földes	143	Szék György/ Szász Susanna/ of földes	Kovács Sándor	Borbála Bábel	
28	Aug. Szept. 15	15	28			János/Anna/ Baetha Bábel/ of földes		Baetha Sándor/ Bábel/ of földes	Kovács Sándor	Borbála Bábel	+1879. 11.
29	Szeptember 20	20	28			János/Anna/ Szék György/ of földes	13	Udvardy Sándor/ Pap Anna/ of földes	Kovács Sándor	Szék György/ Anna	Édes anyja 1879. 9. 11. napján született 200/90
30	Szept. Okt. 27	27	2			Szék György/ Baetha Anna/ of földes	247	Szék György/ Baetha Susanna/ of földes	Kovács Sándor	Leontine	+1879. Okt. 25
31	Okt. 11	11	19			Udvardy Sándor/ Szék Bábel/ of földes	252	Károly György/ Szék Anna/ of földes	Kovács Sándor	Leontine	
32	Szeptember 2	2	9			Szék György/ Károly György/ of földes	16	Horváth György/ Szék Anna/ of földes	Kovács Sándor	Leontine	

Note: The columns contain the following information: the year, month, day for the birth and baptism, the child's name (different table for boys and girls), legitimacy, the name of the parents and their social statute, the house number, name, surname and statute for godparents, the name of the priest, the name of the midwife, observations.

Marriages registers (see Figure 3) accounted numerous actors involved in the event, such as bride and groom, their parents, and godparents (who could be multiple pairs). All need fields for names, civil status, denomination, age, occupation etc. Therefore, the information from this type of source resulted in the largest number of fields in the database.

The Deaths registers (see Figure 4) contained information such as the name of the deceased, denomination, marital status, occupation, residence, birth date, death and burial dates, cause of death, information about parents or spouse, as well as the priest who registered the event. Some registers included more information, such as the legitimacy status if the deceased was a child, the date of the death certificate or mentioned different relatives of the deceased.

Information from the Betrothals registers, which are particular to Orthodox and Greek Catholic Churches, was also included. Betrothals describe the future bride and groom, their parents or guardians (names, literacy, denomination). Although they are interesting life course events, betrothals were usually not recorded, so betrothal registers are not available for every marriage register included in the HPDT. Thus, it was decided to include betrothals in the database, but not to use them for standardization and linkage.

Data entry started by filling a datasheet with information about the source itself. The fields describe county, parish, denomination, type of event, language, alphabet, the dates of first and last records. This datasheet also has a field for the stage of processing of the source (transcribed, checked, standardized), which is updated at later stages.

Figure 3 Excerpt from a marriage register, 1887 (Latin)

Protocolul				cununărilor.			
Anul	Luna	Denominația	Numele, prenume, religia, starea, domiciliul, vârsta, locul nașterii și al locului morții	Numele și prenume, starea, domiciliul, vârsta, locul nașterii și al locului morții	Starea și data sau a treia căsătorie	Când s-au făcut cununile?	Când s-au făcut înscrisurile sau înscrisurile?
1887	Januarie	16	Vasile Andrei grădinar, jura de 24 ani din Harastasi.	Ana Radu grădinar, jura de 21 ani din Harastasi.	1.a	San vestit in 6. 11. 1887	San facut in 6. 11. 1887
1887	Maie	10	George Szigeti grădinar, jura de 24 ani din Harastasi.	Sic. Chiszu grădinar, jura de 21 ani din Harastasi.	1.a	San vestit in 25. April 1887	San facut in 25. April 1887
1887	noiembrie	8	Ioan e. Lăscu grădinar, jura de 24 ani din Harastasi.	Ana Stancu grădinar, jura de 23 ani din Harastasi.	2.a	San vestit in 25. 26. octob. 1887	San facut in 24. octob. 1887
1887	ianuarie	19	George Turban grădinar, jura de 23 ani din Harastasi.	Isuara Dorotea grădinar, jura de 21 ani din Harastasi.	1.a	San vestit in 18. 19. 1887	San facut in 18. 19. 1887
1887	Septembrie	4	George Moldoveanu grădinar, jura de 24 ani din Harastasi.	Olga Turban grădinar, jura de 21 ani din Harastasi.	1.a	San vestit in 28. 29. August 1887	San facut in 27. August 1887
1887	ianuarie	5	Stelu Varga tara grădinar, jura de 25 ani din Harastasi.	Maria Chiszu, jura de 21 ani din Harastasi.	1.a	San vestit in 6. 8. 1887	San facut in 6. 8. 1887
1887	ianuarie	25	Antoniu Măruș grădinar, jura de 25 ani din Harastasi.	Isuara Chiszu grădinar, jura de 21 ani din Harastasi.	1.a	San vestit in 6. 9. 1887	San facut in 6. 9. 1887
1887	ianuarie	29	Ioan Lăscu grădinar, jura de 24 ani din Harastasi.	Maria Rotaru grădinar, jura de 21 ani din Harastasi.	1.a	San vestit in 15. 27. 29. ianuarie 1887	San facut in 14. ianuarie 1887
1887	ianuarie	28	Ioan Stancu grădinar, jura de 24 ani din Harastasi.	Anna Măruș grădinar, jura de 21 ani din Harastasi.	1.a	San vestit in 7. 12. 1887	San facut in 6. ianuarie 1887
1887	ianuarie	21	Ioan Stancu grădinar, jura de 24 ani din Harastasi.	Maria Stancu grădinar, jura de 21 ani din Harastasi.	1.a	San vestit in 14. 21. 28. ianuarie 1887	San facut in 13. ianuarie 1887

Note: The columns contain the following information: name of the deceased, denomination, marital status, occupation, residence, birth date, death and burial dates, cause of death, information about parents and spouse.

Figure 4 Excerpt from a death register, 1860 (Cyrillic)

ПРОТОКОЛЪ МОРЧАЕВЪ							ПРОТОКОЛЪ МОРЧАЕВЪ				
Годъ	Мѣсяцъ	Дни	Имя	Вѣкъ	Причина	Мѣсто погребенія	Имя погребеннаго	Вѣкъ	Причина	Мѣсто погребенія	
1860	Январь	10	Иванъ Ивановъ	60	Старость	Въ церкви	Иванъ Ивановъ	60	Старость	Въ церкви	
1860	Февраль	27	Петръ Петровъ	50	Старость	Въ церкви	Петръ Петровъ	50	Старость	Въ церкви	
1860	Мартъ	26	Иванъ Ивановъ	40	Старость	Въ церкви	Иванъ Ивановъ	40	Старость	Въ церкви	
1860	Апрель	28	Петръ Петровъ	50	Старость	Въ церкви	Петръ Петровъ	50	Старость	Въ церкви	
1860	Май	10	Иванъ Ивановъ	60	Старость	Въ церкви	Иванъ Ивановъ	60	Старость	Въ церкви	
1860	Июнь	20	Петръ Петровъ	50	Старость	Въ церкви	Петръ Петровъ	50	Старость	Въ церкви	
1860	Июль	20	Иванъ Ивановъ	60	Старость	Въ церкви	Иванъ Ивановъ	60	Старость	Въ церкви	
1860	Августъ	20	Петръ Петровъ	50	Старость	Въ церкви	Петръ Петровъ	50	Старость	Въ церкви	
1860	Сентябрь	20	Иванъ Ивановъ	60	Старость	Въ церкви	Иванъ Ивановъ	60	Старость	Въ церкви	
1860	Октябрь	20	Петръ Петровъ	50	Старость	Въ церкви	Петръ Петровъ	50	Старость	Въ церкви	
1860	Ноябрь	20	Иванъ Ивановъ	60	Старость	Въ церкви	Иванъ Ивановъ	60	Старость	Въ церкви	
1860	Декабрь	20	Петръ Петровъ	50	Старость	Въ церкви	Петръ Петровъ	50	Старость	Въ церкви	

Note: The columns contain the following information: year, month, day of the death, burial date, name of the deceased, denomination, place of burial, name of the priest, the age of the deceased, cause of death, observations.

2.2 SAMPLE STRUCTURE

Since registers covering the project period are kept in the county departments of the National Archive, photographing documents was an important activity. Analysis of the sources, including review of inventories of archival collections was one of the first tasks of the research team, followed by the development of a methodology for selecting micro-areas included in the database. The target was to cover between 5 and 10% of the Transylvanian population from the period 1850–1914. By the end of the project, 7% was included.

The sampling framework was built on several prerequisites (Crăciun, Holom, Popovici, 2015). The first one was continuity of the sources: the parish records must have information for at least 50 consecutive years. Then, the sources had to include the ethnic and denominational composition of Transylvania, ensuring balance with regard to ethnicity and denomination, while taking into account the geographic, ethnocultural, and historical unity/homogeneity of certain regions. An important selection criterion was developed to balance the countryside and settlements with an integrative role, like urban and semi-urban centres. The resulting sample is divided into 12 micro zones which consistently cover almost 7% of the historical population of Transylvania. For an extensive discussion of the sample selection procedures, see Crăciun et al. (2015).

Teams of researchers went to 10 Transylvanian county branches of the National Archive of Romania, gathering the sources and assuring their primary organization (photographing, processing and cataloguing the images). Gathering sources has been an ongoing activity, and there are currently around 500,000 images in the repository of sources. It should be mentioned that no complete and accurate catalogue of these sources in the County Archives Services existed until recently. One of our tasks was to gather all information on existing parish registers and build an electronic catalogue. We estimate that Transylvanian archives hold over 30,000 parish registers, covering seven major denominations, written in five main languages and three alphabets.

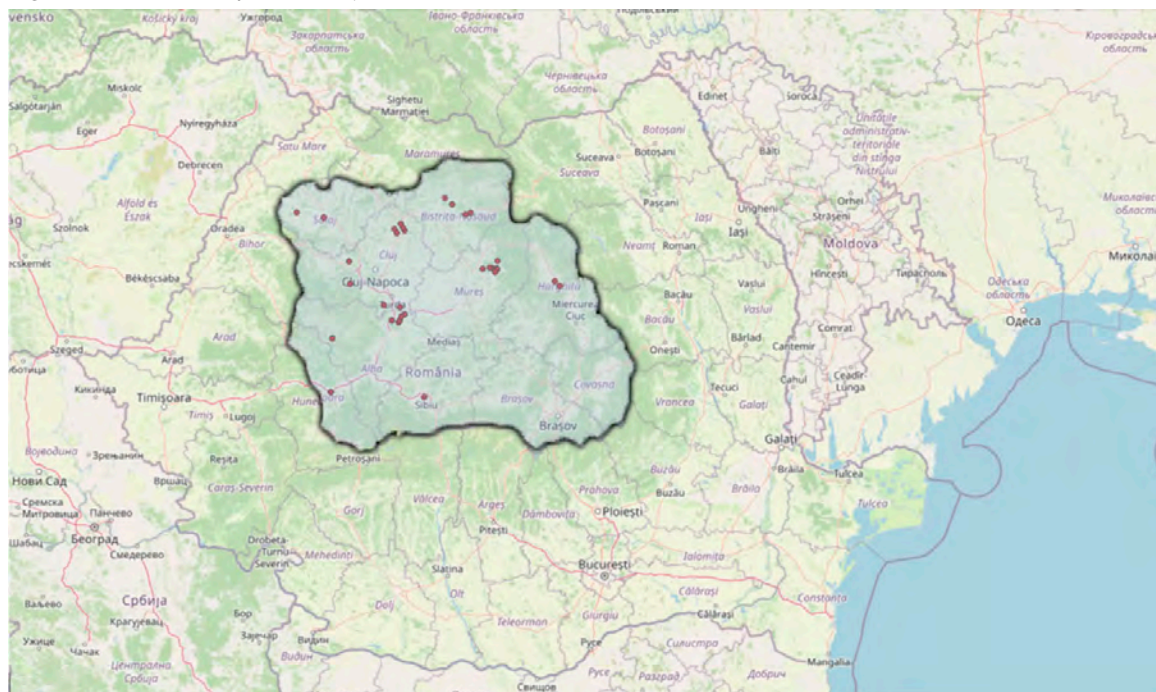
Confronting the sources revealed challenges for the team. Registers from the first locality included in the sample (Călărași, Cluj county) complied with all of the initial prerequisites, but the registers proved to have very sparse information. Thus, we decided to adapt the structure of the sample by adding the richness of information in the registers to the initial sampling criteria. A decisive argument for our approach was the linkage process performed on the first locality, which encountered numerous difficulties because of the lack of information in the sources.

From November 2014 to February 2021, 25 localities were included in HPDT, totalling 165 parish registers. Table 1 provides an overview of the current status of the database containing 141,038 events.

Table 1 *Number of locations and events, included in the HPDT database, February 2021*

Denomination	BIRTHS		DEATHS		MARRIAGES		BETHROTALS		Total	
	Locations	Events	Locations	Events	Locations	Events	Locations	Events	Locations	Events
Civil registration			1	1,337					1	1,337
Jewish	2	1,660	2	397	2	377			6	2,434
Roman Catholic	5	5,995	5	6,292	4	1,927			14	14,214
Greek Catholic	15	21,503	14	15,974	14	5,817	1	188	44	43,482
Orthodox	8	12,894	7	8,433	7	2,630	2	447	24	24,404
Lutheran ¹	1	17,241	1	12,696	1	5,648			3	35,585
Calvinist (Reformed)	8	8,807	8	7,476	9	3,299			25	19,582
Total	39	68,100	38	52,605	37	19,698	3	635	117	141,038

1 This represents the special case of Sibiu, a town from the southern part the selected area, one of the Transylvanian places inhabited by Saxons. Sibiu served as a case study for a doctoral thesis defended at University of Regensburg in 2021 and the HPDT team gladly accepted to host the data and to provide the database architecture. The data are particularly interesting since Sibiu is one of the big towns of Transylvania, with a population structure more differentiated in terms of education, occupations, intergenerational changes, which will allow complex analyses into the future.

Figure 5 *Map of Transylvania and the localities included in HPDT*

In February 2021 there were 570,036 individuals recorded in HPDT (360,000 in April 2017, at the end of the project). Although the standard period of the database is 1850–1914, due to the necessity of preserving the unity of the archival collection, some localities contain data recorded before 1850 or after 1914. For instance, if the locality preserved records starting from 1780, the project team decided to incorporate the years before 1850 to maintain the integrity of the source and for research purposes. In some situations, the registers did not cover the entire time span of the project or the records did not cover all denominations (the Roman-Catholics and the Reformed started to keep evidence sooner than the Orthodox, and their registers are much better preserved).

2.3 GENERAL PRINCIPLES OF DATA ENTRY

The database building process started from the protocol provided by Mandemakers and Dillon (2004). Information was transcribed literally into the database, avoiding any form of abbreviation or standardization. A detailed instruction manual for transcription was created and continuously updated during the process (Boloan et al., 2019). The data-entry operators were trained with this manual, and they had permanent access to it. However, the variety in information contained in the church registers (multiple languages, different alphabets, and different denominations) required a significant number of adjustments and adaptations in order to conform to international standards of database creation. This meant successive adaptations of the data entry manual and of the database architecture as new particularities of the sources unfolded.

As a general principle of data entry, information from the source was literally transcribed in forms constructed for each type of event (birth, marriage, engagement, death). The data-input forms have various types of fields: text fields, dropdown lists and checkboxes. The data entry operator transcribed information exactly as it appears in the source into text fields, including the errors or misspellings, if any. In this way, the original text of the source was preserved. Such fields included first and last name, nickname, occupation, cause of death etc.

Since the team was reduced in size, the time pressure was high, and some information was repetitive, a decision was made for a preliminary basic standardization of a few fields, in order to avoid redundancy and errors of transcription during the data-entry process. Later standardization would have implied a lot of resources to code and integrate these data into the database structure. Denomination, gender and literacy were standardized for each individual in the event (parents of the baptized child, of the bride and groom, of the deceased, as well as witnesses and godparents), legitimacy for the baptized child, and marital status of the bride and groom or the deceased. Dropdown lists with predetermined values were used for these fields from which the appropriate element must be selected in accordance

with the information in the register. For example, the Denomination list includes "Christian", "Jew", "Orthodox", "Greek-Catholic", "Roman-Catholic", "Lutheran".

A similar approach was used for priests and midwives. The same priest could have officiated over 2,000 events (baptisms, marriages and burials); a midwife could have assisted tens of births. It was more efficient to gather all priests present in the sources into one dropdown list, which was updated as new priests appeared in the registers. The same was done with the midwives.

A third type of field used in the data entry forms was checkboxes in which the appropriate value must be ticked, such as stillborn, multiple birth or Julian calendar. The same approach was used to add Observations/Comments, where a ticked box opens a text field to be filled in.

The data entry forms have fields for all the information likely to be recorded in registers, even though the sources differ greatly in the quality and quantity of data. For example, the forms have fields for the occupation of all possible roles (the parents of the baptized, the brides and grooms, the godparents and so on), but many registers lack this information.

2.4 THE DATABASE

The original database consisted of tables for each of the vital events found in the parish registers, reflecting our source-oriented approach. Over time, the database and data entry forms evolved as we gained more experience and encountered more complex sources. Features like drop-down lists were added to standardize repetitive information and speed data entry. The database was "normalized" by adding tables for witnesses, godparents, and other participants who appeared in varying numbers. Even when an event is stored in multiple tables, the original document can be reconstituted digitally using keys that link tables to each other. The following section describes phases in the restructuring of the database, which are shown in Figures 6–9 in the Appendix.

2.4.1 DEVELOPMENT AND BASIC STRUCTURE

The architecture of the source database needed to combine different parish registers from different denominations into a single structure. It was structured around the main vital events (births, betrothals, marriages, deaths) and included information not only about individuals but also about the event itself. The parish registers from Transylvania contain very diverse information, and each denomination has their own system of recording information with different column headings that changed over time. The database had to accommodate all the different fields from the parish registers so the number of columns for each main table of the database increased to a very large number. However, every type of information was not found in all the registers, and columns would have been empty in more than 70% of records. The information recorded by the priests was sometimes very basic, reduced to the names and possibly a date and place, even though the registers had columns for more information. Tables with more than 200 columns were difficult to manage and had performance issues in retrieving data. Moreover, the empty columns were unnecessarily increasing the size of the database. The solution was to normalize the database and organize the information in linked tables.

In the first version *hpdv1* built in 2014, each data entry form describing a main event included all the different fields found in the parish registers from all denominations attached to a table in the database. Information was structured in 17 tables and 475 columns (see Appendix, Figure 6). Although the database followed the source-oriented method, a decision was made to use standardized forms for repetitive information throughout the registers. As we already mentioned, this included denomination, gender, legitimacy and all information regarding priests and midwives. Priests and midwives were identified by a unique combination of name and place of service.

When the richness of the information in the registers became a main criterion for inclusion in the sample (see Section 2.2), new fields were added in the data entry forms and accommodated in the database as new columns in the underlying tables (for example second occupation for all participants in the events, relation between godparents, information about the spouse of the deceased, ethnicity). Additional data entry forms were added for information about confirmation, converts, and name changes, as this information was found in some of the parish registers. Thus, the second version *hpdv2*, implemented by the end of 2015, contained 29 tables and 690 columns (see Appendix, Figure 7). Many of these tables are relational and allow for many-to-many relationships, but they are not included in the graphical interface shown in Figure 7.

In 2016 a third version was necessary, *hpdt_v3*. During the data-entry, new fields emerged, and the database needed to be restructured. The main change was moving the godparents sections from Births and Marriages to a new table. These event tables could accommodate one pair of godparents with information that could entail 15 columns. Further data entry encountered events with multiple pairs of godparents present, ranging from 2 to 5 pairs. Accommodating all of them would have unduly increased the number of columns in the event table, given that a majority of the events recorded had only a single pair. In order to avoid repeating fields the database was further normalized by creating a new table — Godparents — with a corresponding data entry form for the operator. Every pair of godparents was linked by the marriage or birth ID. The complete event was reconstituted in the detailed view of the user interface, so nothing from the original recording of the event is lost. *Hpdt_v3* had 36 tables and 700 columns (see Appendix, Figure 8). So, the detailed view of the data entry form also contains not only the data that has to be entered but also all relevant information from the related tables.

In 2020, normalization was applied to the marriage witnesses, leading to the fourth version, *hpdt_v4*. The data entry form for marriage witnesses allowed up to seven individuals to be recorded, but events including seven witnesses were extremely rare, so many columns remained empty. Another new addition was the table Cause of Death. The cause of death was still recorded in the original language and exactly as it appears in the registers, but all values were written into a dedicated table along with standardized fields on the data entry form. The result was a mix between standardization and retaining the original text of the written source. This change was needed to speed up the process of standardization and to eliminate the data redundancy, a procedure done by the data entry operator. A third addition was the table Death Relatives for relatives present at the Burial event, which used an approach similar to the ones applied to godparents and marriage witnesses to cope with extremely heterogeneous information. With all the newly added tables, *hpdt_v4* has 43 table and 829 columns (see Appendix, Figure 9).

Even though information was separated in different tables in the database architecture, the whole source can be reconstituted digitally to display a copy of the original written record. Everything is linked with the table Sources, in which the original sources are recorded. The user can retrieve the archival code of the registry, the page where the event was mentioned, the languages used and the location. When accessing the information related to the source, the number of records from that particular source is displayed for the user to see how large the source was. The database has a user interface accessible by login.

To summarize, the database is mainly built around four major tables corresponding to the types of vital events recorded in the parish registers. Each table details a single type of religious event: baptism (Births), engagement (Betrothals), marriage (Marriages) and burial (Deaths). Within the tables each row describes a single event. Several other tables include specific persons, such as godparents, witnesses, relatives present at the main events. In addition to the main tables, complementary tables provide values for the dropdown lists that standardize the database. Tables are linked to each other by keys allowing the information to be combined. Tables providing objects for value lists are: Source, Priests, Midwives, Converts, Confirmation, Name change, Denominations, Ethnicities, Countries, Genders, Legitimacies, Dispensations, etc. All events from the main tables are linked to the original written source, and the whole source can be reconstituted digitally.

It is clear that the database is still evolving. If new types of information are found during the transcription process, new fields are added to accommodate the original source. Search filters are provided for researchers in order to find and sort records according to their needs. Everything is interrelated, making the extraction of the necessary information an easy task.

2.4.2 INTEGRATED DATABASE: STANDARDIZATION

The second component of the research database is the standard database, which is a copy of the central source database with some additional fields. In the structural metadata of three main tables (Births, Marriages, Deaths) new columns were added for standard names, standard age, and standard places. The original source values were also preserved alongside the standard versions. Values already standardised from the tables for denominations, civil status, legitimacy, literacy, priests, and midwives were not modified. In addition to the tables originating from the source database, several tables have been created to aid in the standardization and record linkage processes. One of the most important tables was the Names table. Standardisation has been applied to the first name, last name and nickname. As we stated before, one of the major characteristics of Transylvania was the use of several languages

in the parish registers: Romanian, Hungarian, German, and Latin. Romanian and Hungarian names were processed each in their respective language. Romanian names were assigned for the Orthodox and Greek-Catholics, while the Catholics and Calvinists were associated with Hungarian names. The presence of Hungarian names in the Romanian registers and vice versa were considered acceptable losses. Names like "Catalina" were standardised to "Cătălina" (Romanian) or "Katalin" (Hungarian). Several names with multiple variants, like diminutives, were standardised to the most common one.

In addition to names, other variables were standardized, such as age, location, occupations and causes of death. Age was standardized and divided into several columns for years, months, days. Locations were coded and standardized according to the local administrative unit. We adopted the encoding method that is largely used within the European Union (<http://ec.europa.eu/eurostat/web/nuts/local-administrative-units>).

The sources included in HPDT provide information on the occupations of individuals with different roles in vital events: mother, father, grandparents, and godparents of the baptised child; groom, bride, parents, godparents, and witnesses in a marriage; the deceased person, parents, and spouse. Following our participation in a workshop on IDS for Population Registers in Lund, Sweden, September 2015, the HPDT core team decided to begin the process of coding Transylvanian historical occupations using the Historical International Standard Classification of Occupation (HISCO) scheme (van Leeuwen, Maas, & Miles, 2002). This scheme, used worldwide, is a necessary step not only to do comparative research but also to standardise the enormous heterogeneity in terms of languages and alphabets in the Transylvanian database. Thus, occupations such as "miller" may appear written in Latin as "molendinarius" or "molitor", as "molnár" in Hungarian, "Müller" in German, or "morar" in Romanian. By using the HISCO scheme, all these linguistic variations are brought together under one single code: 77120.

The coding process of occupations followed several steps: 1. Standardizing the names of the occupations, proofreading and reviewing, the extension of abbreviations; 2. The translation of the occupation names into English; 3. Assigning HISCO codes; 4. Assigning HISCLASS codes; 5. Assigning SOCPO codes. Further, specific historical class or social status schemes, such as HISCLASS (van Leeuwen & Maas, 2011) or SOCPO (Van de Putte & Miles, 2005) were built. For instance, the "miller" with a 77120 code in HISCO, becomes 7 (medium skilled workers) in HISCLASS and 2 (semi-skilled) in the SOCPO scheme. The process of transformation of HISCO codes from HPDT in HISCLASS and SOCPO is undergoing, but these two schemas have already been used in some analyses carried out by members of the Centre for Population Studies despite some misregistration of occupational titles in the parish registers from Transylvania (Holom, Sorescu-Iudean, & Hărăguș, 2018, p. 339).

As part of standardization in HPDT, causes of death have been subjected to linguistic equivalence and coding designed to facilitate the use of information for future demographic research. The coding process involved several steps: 1. Standardization of the causes of death, proofreading, reviewing and completion of abbreviations; 2. English translation (the modern correspondent in English language); 3. Assigning of codes in the International Statistical Classification of Diseases and Related Health Problems-ICD-10.; 4. Assigning of codes in the Historical Causes of Death-HCD (HCD) system (World Health Organization [WHO], 2016; Holom & Hegedűs, 2021).²

The process of standardization has been essential not only because registration was in Hungarian, Latin or German, but also because sometimes causes of death appeared in parish registers using colloquial language. For example, "gutaütés", "lovit de gută", "apoplexie", "Schlagfluß" are all terms used to designate stroke. The next step was to code each cause of death following the ICD-10 standard to facilitate future analysis, wider accessibility, and the possibility of comparative studies. Coding into ICD-10 proved to be very problematic. Causes of death mentioned in Transylvanian parish registers from the 18th and 19th centuries are often focused on symptoms (e. g. fever, cough), or are extremely vague (e.g. "natural death", "old age" or "3 days long sickness"), since the vast majority of records were not made by established specialists. Therefore, coding on the basis of ICD-10 created significant difficulties, which we want to overcome by developing a special coding system, the HCD, which is more compatible with data on historical populations.

The Historical Causes of Death system (HCD) is structured in eight categories: 1. Infectious diseases; 2. Chronic and acute non-infectious diseases; 3. Diseases originating in the perinatal period; 4. Diseases

2 This part of documenting the process of coding of causes of death is a work in progress, and as such the name and the structure of the new system that we intend to develop may have future changes.

related to pregnancy, childbirth and childbed period; 5. Old age-related diseases; 6. Violent deaths; 7. Symptoms, signs and abnormal findings; 8. Ill-defined and unknown causes of mortality. As such the Hungarian term "gutaütés" received the code I64 in ICD-10 and 2 in HCD system. We consider the HCD system suitable for future analysis of mortality in Transylvania, and it can be used by other scholars studying causes of death.

2.4.3 INTEGRATED DATABASE: THE LINKAGE PROCESS

At this time, record linkage has only been applied to data from the first locality included in HPDT, namely Călărași, Cluj county. The sample consisted of 2,497 births (baptisms), 1,020 marriages and 2,577 deaths and the individual names, age, locations and birth dates were previously standardized. Only individuals with a main role in the event were included, resulting in a sample of 14,311 individuals from the three types of registers. The fields extracted for record linkage included original names, standard names, location (birth place, residence, wedding place), gender, birth year, event year (when the individual is mentioned), wedding year, death year. First name, last name, sex and role in the event were the variables used to link persons appearing in the registers to unique individuals.

We first linked parents to their (multiple) children in the baptism records. If there were multiple children born from the same parents within a certain time interval the record linkage program assigned the same id number in order to reconstitute the family. Then, we linked baptisms records with the parental marriage records by identifying parents from the baptism records with the bride and groom from a marriage record. Then we linked deaths records as well, identifying parents from the baptism records with their death record. Through a similar process, we identified children in the baptism records as spouses in marriage records and/or as deceased, in deaths records. Difficulties were encountered in every stage, and the linkage process turned into a semi-automatic one.

The software used for record linkage is based on Jaro-Winkler similarities between names. It was developed at the Arctic University of Norway, Tromsø and adapted to the realities of Transylvania. The program uses three levels of the result score: level 1 – score 0.96 (most probable match), level 2 – score 0.90 (probable match), level 3 – score 0.80 (possible match). When using the first level the computer writes the link (gives the matched persons the same ID) automatically into the database. Level 2 and 3 have to be checked by the user before the computer enters the established identification into the database. Only the standard names have been used because the original names had great variability in spelling.

Ethnic and denominational diversity of Transylvania created several challenges. If a person was recorded both in a Romanian and in a Hungarian register with names translated into the respective languages (e.g. "Ioan"/"János"), it was impossible for the automated linkage software to identify he/she as the same person. In this case, only manual linkage could be used because the Romanian and Hungarian versions of a name get low similarity scores.

Naming practices for the female population are also a problem. While Hungarian women kept their maiden name after marriage³, there was no apparent rule for last names after marriage among Romanian women. Sometimes they were mentioned with their maiden name, other times by their husband's last name, but most of the time they lacked last names altogether. Several variables, like maiden name and role in the event, were constructed in order to adjust the linkage process to this historical reality (Wisselgren, Edvinsson, Berggren, & Larsson, 2014).

However, the biggest problem for accurate linkage was lack of information in the sources, which led to a system of variables created through inference. A semi-automatic linkage was developed for incomplete or heterogeneous information. A series of stored procedures with different query conditions and a Jaro-Winkler function extracted files with possible connections that were checked manually.

The need for homogenization of information for all confessions, missing information, and the construction of multiple supplementary variables were challenges for the linkage process in Transylvania. The results obtained under these circumstances — 73% success rate for linking parents and children, 29% for linking baptism and marriage records and 21% for linking baptism and death records — were a solid argument in favour of the choice to select sources with richer information. The database is still in development, data-

3 The Hungarian practice of adding the "né" suffix after the husband's name in order to underline the marital status of the wife is well documented (Fercsik, 2010).

entry is a continuing process, and standardization and record linkage are still in an early phase. The next logical steps are to advance with these processes, while accommodating them to Transylvanian realities.

3 HPDT AS AN INSTRUMENT FOR RESEARCH

The HPDT longitudinal database was created for research on the population of Transylvania in the 19th and 20th centuries. The database opens new directions of research in areas such as history, demographics, sociology, economy, linguistics, and medical history. The period of time covered by the HPDT represents a crucial era for the study of fertility decline, urbanization, household composition, occupational structure, gender equality. By providing individual-level data, the database allows for statistical analyses with advanced methods on questions that have received very little investigation from a historical perspective. In this section we provide an overview of the main published analyses conducted with data extracted from the HPDT.

An analysis of infant mortality in rural parishes in Transylvania in the second half of the 19th century indicated that almost half of deaths occurred in the first month of life. This disastrous reality was a consequence of the poor conditions of sanitation during the period of pregnancy and the moment of child-birth. Children strong enough to survive the neonatal period mostly died from epidemic diseases, and male infants were the most vulnerable (Coroian, 2017). An investigation of the seasonality of mortality in three Transylvanian settlements between 1887 and 1912 highlighted higher levels during spring and winter, the most vulnerable groups being infants and children (Coroian, 2016).

A working sample of 6,719 adults who died after age 24 has been analysed taking into account both environmental and individual conditions, such as locality type, period, marital status and socioeconomic status. The findings indicated that adults in open localities undergoing industrialization were more prone to premature death, than those living in peripheral, agricultural localities. Between 1850 and 1880, adult mortality was influenced to a greater extent by environmental and epidemiological crisis, but differences were due to economic development and working activities between 1881 and 1914. The main beneficiaries of investments made after 1881 in industry, technology, public sanitation and health care were males (men in agricultural occupations, men employed as semiskilled workers, and unmarried men). Marriage had a protective effect on men, but not on women. After the 1880s, survival prospects improved for both males and females (Holom, Hărăguș, & Bolovan, 2021).

Previous research on age at first marriage in Transylvania focused on nuptial realities in small, isolated villages. With data from the HPDT, we could construct a consistent and coherent sample to consider more explanatory factors. Holom et al. (2018) analysed several factors that influenced the age at first marriage, such as denomination, migration background, and socio-occupational status, as well as broader determinants, such as the time frame and the level of development achieved by settlements under study. Some of the findings were in accord with other areas in Europe, such as the tendency of Roman Catholics to marry later, or the postponement of marriage among men and women with a migrant background. The data indicated that Calvinist women and self-employed men tended to marry later, while both men and women in less developed areas married earlier. The article also took into consideration the interaction between individual factors and broader realities. It found that the level of development of localities was in many cases more important than individual co-variables in determining constraints and opportunities on the marriage market.

Combining HPDT data with information from enrolment records and cadastral registers, Botoș (2019) studied social mobility and the role of education in the Gurghiu Valley, located in the eastern part of Transylvania. Occupational titles were coded into HISCO, and later into HISCLASS 5 (Mandemakers et al., 2013). Her findings indicate that society in the Gurghiu Valley was chiefly agrarian and immobile, and only a small number of people were able to climb the social ladder through education.

The second half of the 19th century witnessed the medicalization of childbirth in Europe and in Transylvania as well. Dumănescu and Eppel (2019) approached the “medicalization” process in both its meanings: the professionalization of healers and the spread of medical care. The paper describes the training of midwives in a society on the periphery of Austro-Hungarian society and the introduction of modern care into one of the most intimate moments of a woman's private life.

Interest in midwifery in Transylvania came from the abundance of midwives in the parish registers included into the Historical Population Database of Transylvania. Over a period of less than 50 years, thousands of women performed deliveries in villages and had their names written in the "midwife" column in the registers. This discovery raised questions: Who were these midwives? Were they really midwives? Were they trained, skilled midwives or simply handy women who helped deliver their neighbours' babies? The preliminary results of Dumănescu and Bolovan (2021a) confirm that the medicalization of childbirth at the turn of the 20th century was not nearly as widespread as one might expect from official statistics. Over 80% of the women reported in our sample were handywomen, rather than certified midwives. One could also conclude that the medicalization in all its aspects, including childbirth, went hand in hand with the processes of modernisation and industrialisation.

A study concerning women and their married names in Transylvania in the second half of 19th century Dumănescu and Bolovan (2021b), based on a sample which included 29,000 baptisms, 3,982 weddings, and 6,592 events of death, revealed that a marriage contract was not automatically followed by a change in name and that a married woman was still recognized by her maiden name in subsequent documents.

A study of so-called "necronymic" names (Mărza, 2017) aimed to reconstruct certain relational patterns over several generations of the same family. The article begins from the assumption that the naming of children was not random. This implies that the recurrence of certain names across three or four generations could be an indication of the type and the quality of the constructed links in a family. The practice of "necronymic naming" (naming a child after a deceased sibling) was considered a method of strengthening and perpetuating the fabric of the family. Working on the data from two localities in the HPDT, this article highlighted the limitations of the vital registers for accurately reconstituting all families on both the maternal and the paternal lines. Even when the standardization of HPDT will be complete, the lack of certain essential data (such as the mother's last name in many birth registrations) sometimes limits the reconstitution of families to only one or two generations.

4 CONCLUSIONS

From the beginning, HPDT was intended to become an instrument dedicated to researchers with an interest in the past of Transylvania. Designing and building a database, especially when using a source-oriented approach was a difficult task, and HPDT had to cope with complex sources providing heterogeneous information. The intermediate versions of HPDT show how it has continuously evolved and adapted to meet these challenges, ensuring its value for a wide range of research questions and methodologies.

The source-oriented approach preserved the original documents in their entirety, but stored them in an organized form. Standardization of information was the next step for integrating Transylvanian data into international research. Standardization occurred in two steps. The initial step standardized and codified repetitive data, such as denomination, ethnicity, marital status, as well as priests and midwives who appeared frequently in the sources, to eliminate redundancy and reduce transcription errors during the data-entry process. A second round of standardization was applied at the end of data input for selected communities. This time all names, occupations, locations, and causes of death were standardized and codified according to international standards for historical databases. Standardization allowed a partly automated linkage process, using software based on the Jaro-Winkler distance for similarities. However, the lack of uniformity in data from the various Transylvanian denominations is still an impediment to increasing the automation of the linkage process, which will need to be addressed in the future.

The real impact of HPDT can be best assessed from the diversity of studies already published or in progress, which reflects its increased importance for historical, demographic, sociological, and anthropological research. HPDT is continuously adapting to meet the needs of researchers applying micro-level quantitative perspectives to the study of the historical past in Transylvania. Their studies will be better integrated into universal knowledge of the past, because alignment with international standards of databases for historical data makes them suitable for comparisons across national borders.

REFERENCES

- Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies*, 1, 1–26. doi: [10.51964/hlcs9290](https://doi.org/10.51964/hlcs9290)
- Bolovan, I., Crăciun, B., Covaci, D., Dumănescu, L., Holom, E.-C., Mârza, D., & Lumezeanu, A. C. (2019). Historical Population Database of Transylvania. A database manual. *Studia Universitatis Babeș-Bolyai Digitalia*, 64(1), 9–84. doi: [10.24193/subbdigitalia.2019.1.1](https://doi.org/10.24193/subbdigitalia.2019.1.1)
- Botoș, R. (2019). Education as a vehicle for social mobility in the 19th century in Transylvania. A comparative view on Romanians and Hungarians in the Gurghiu valley. *Romanian Journal of Population Studies*, 13(1), 29–46. doi: [10.24193/RJPS.2019.1.02](https://doi.org/10.24193/RJPS.2019.1.02)
- Coroian, I. G. (2016). The seasonality of mortality in three Transylvanian settlements in the second half of the 19th century. *Romanian Journal of Population Studies*, 10(1), 19–35.
- Coroian, I. G. (2017). Infant mortality in rural Transylvania: A case study on four parishes in the second half of the 19th century. *The Romanian Journal of Modern History*, 8(1–2), 5–18.
- Crăciun, B., Holom, E. C., & Popovici, V. (2015). Historical Population Database on Transylvania: Methodology employed in the selection of settlements and micro zones of interest. *Romanian Journal of Population Studies*, 9(1), 17–31.
- Dumănescu, L., & Eppel, M. (2019). The politics of birth in a composite state: Midwives in Transylvania (19th–20th century). *Romanian Journal of Population Studies*, 13(1), 7–27. doi: [10.24193/RJPS.2019.1.01](https://doi.org/10.24193/RJPS.2019.1.01)
- Dumănescu, L., & Bolovan, I. (2021a). Medicalisation of birth in Transylvania in the second half of the 19th century. A subject to be investigated. *Historical Life Course Studies*, 10(3), 91–95. doi: [10.51964/hlcs9574](https://doi.org/10.51964/hlcs9574)
- Dumănescu, L., & Bolovan, I. (2021b). 'From the cradle to the grave I am my father's daughter!' Women and their married names in Transylvania in the second half of 19th century. *The History of the Family*, 26(3), 466–481. doi: [10.1080/1081602X.2021.1933126](https://doi.org/10.1080/1081602X.2021.1933126)
- Fercsik, E. (2010). The traditional and modern forms of Hungarian female matrimonial names. In M. G. Arcamone, D. Bremer, D. De Camilli & B. Porcelli (Eds.), *Atti del XXII Congresso Internazionale di Scienze Onomastiche Pisa, 28 agosto – 4 settembre 2005* (Vol. IV, Antroponomastica) (pp. 131–140). Pisa: Edizioni Ets. Retrieved from <https://mnytud.arts.unideb.hu/nevtan/informaciok/pisa/fe-a.pdf>
- Holom, E. C., Hărăguș, M., & Bolovan, I. (2021). Socioeconomic and marital status inequalities in longevity: Adult mortality in Transylvania, 1850–1914. *Journal of Interdisciplinary History*, 51(4), 533–564. doi: [10.1162/jinh_a_01627](https://doi.org/10.1162/jinh_a_01627)
- Holom, E. C., & Hegedűs, N. (2021, April). From ICD-10 to a new nosological classification of causes of death in Transylvania, 1850 and 1920. Talk presented at the *Local Population Studies Society Spring Conference*, Local Population Studies Society and the Southampton Centre for Nineteenth-Century Study, University of Southampton, U.K.
- Holom, E. C., Sorescu-ludean, O., & Hărăguș, M. (2018). Beyond the visible pattern: Historical particularities, development, and age at first marriage in Transylvania, 1850–1914. *The History of the Family*, 23(2), 329–358. doi: [10.1080/1081602X.2018.1433702](https://doi.org/10.1080/1081602X.2018.1433702)
- Lumezeanu, A.-C. (2019). *Digital infrastructure for social history. Building historical databases* (Doctoral dissertation). Babeș-Bolyai University, Cluj-Napoca.
- Mandemakers, K., & Dillon, L. (2004). Best practices with large databases on historical populations. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 37(1), 34–38. doi: [10.3200/HMTS.37.1.34-38](https://doi.org/10.3200/HMTS.37.1.34-38)
- Mandemakers, K., Muurling, S., Maas, I., Van de Putte, B., Zijdeman, R. L., Lambert, P. S., van Leeuwen, M. H. D., van Poppel, F. W. A., & Miles, A. (2013). *HSN standardized, HISCO-coded and classified occupational titles, release 2013.01*. Amsterdam: IISG.
- Mârza, D. (2017). Patterns in family relationships in 19th century Transylvania: Data from the Historical Population Database of Transylvania. *Transylvanian Review*, 26(4), 63–70.
- Van de Putte, B., & Miles, A. (2005). A social classification scheme for historical occupational data. *Historical Methods. A Journal of Quantitative and Interdisciplinary History*, 38(2), 61–94. doi: [10.3200/HMTS.38.2.61-94](https://doi.org/10.3200/HMTS.38.2.61-94)
- van Leeuwen, M. H. D., Maas, I., & Miles, A. (2002). *HISCO. Historical International Standard Classification of Occupations*. Leuven: Leuven University Press.
- van Leeuwen, M. H. D., & Maas, I. (2011). *HISCLASS. A Historical International Social Class Scheme*. Leuven: Leuven University Press.

Wisselgren, M. J., Edvinsson, S., Berggren, M., & Larsson, M. (2014). Testing methods of record linkage on Swedish censuses. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 47(3), 138–151. doi: [10.1080/01615440.2014.913967](https://doi.org/10.1080/01615440.2014.913967)

World Health Organization [WHO]. (2016). *ICD-10: International statistical classification of diseases and related health problems, 10th revision, 5th edition* (Vol. 1–3). Retrieved from <https://apps.who.int/iris/handle/10665/246208>

APPENDIX — MODELS OF THE DATABASE

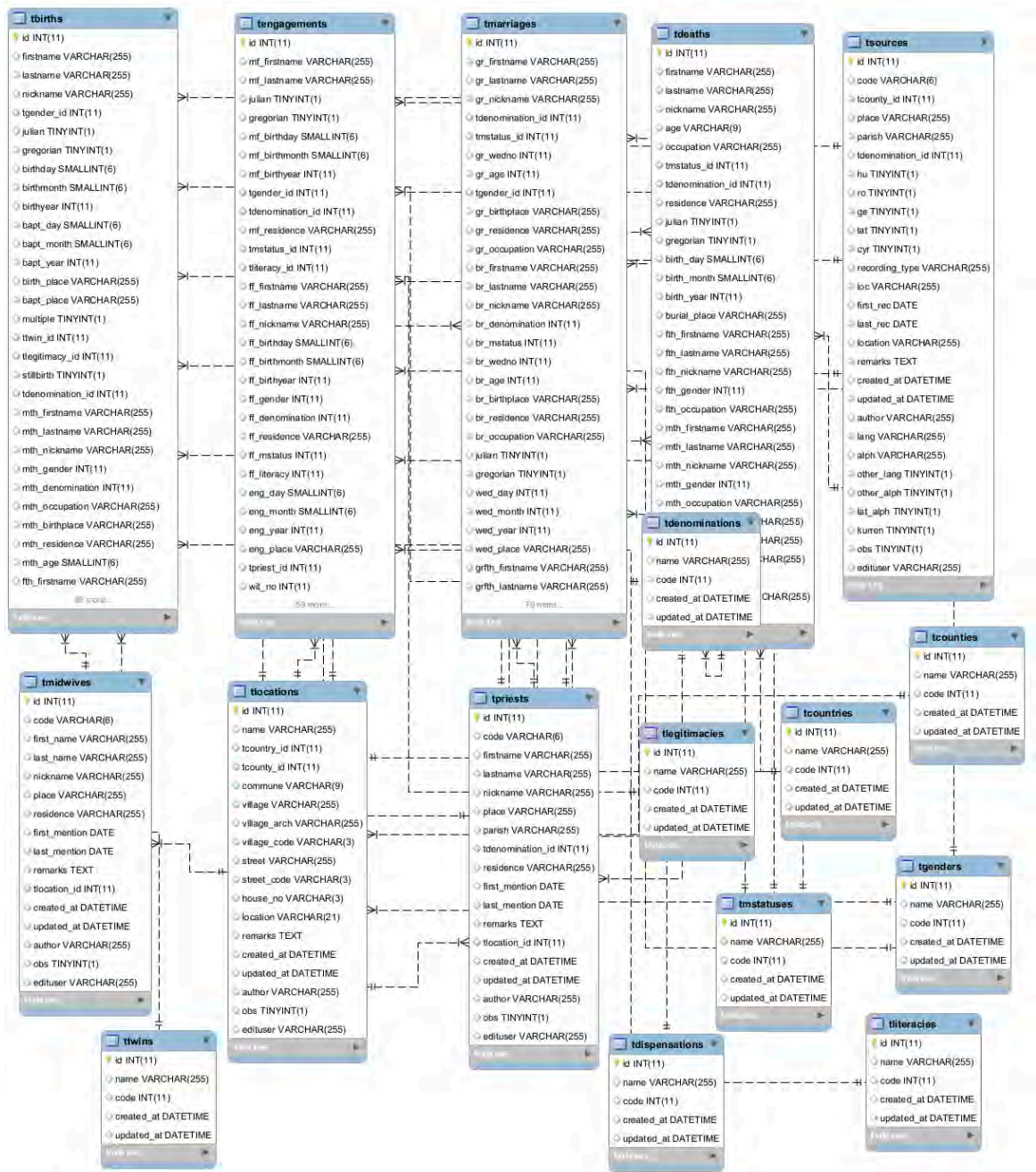


Figure 8 *Model hpdt_v3*

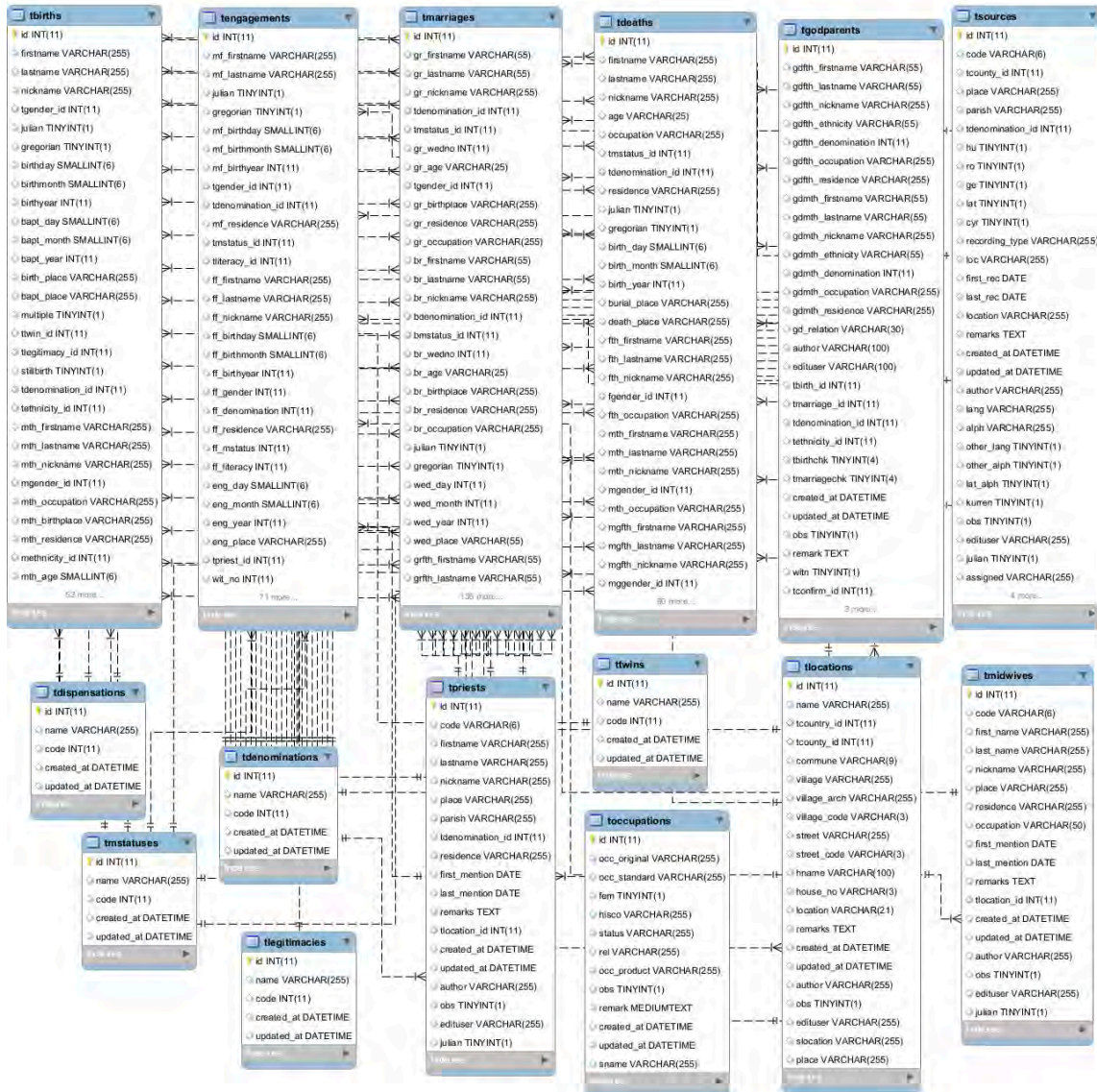
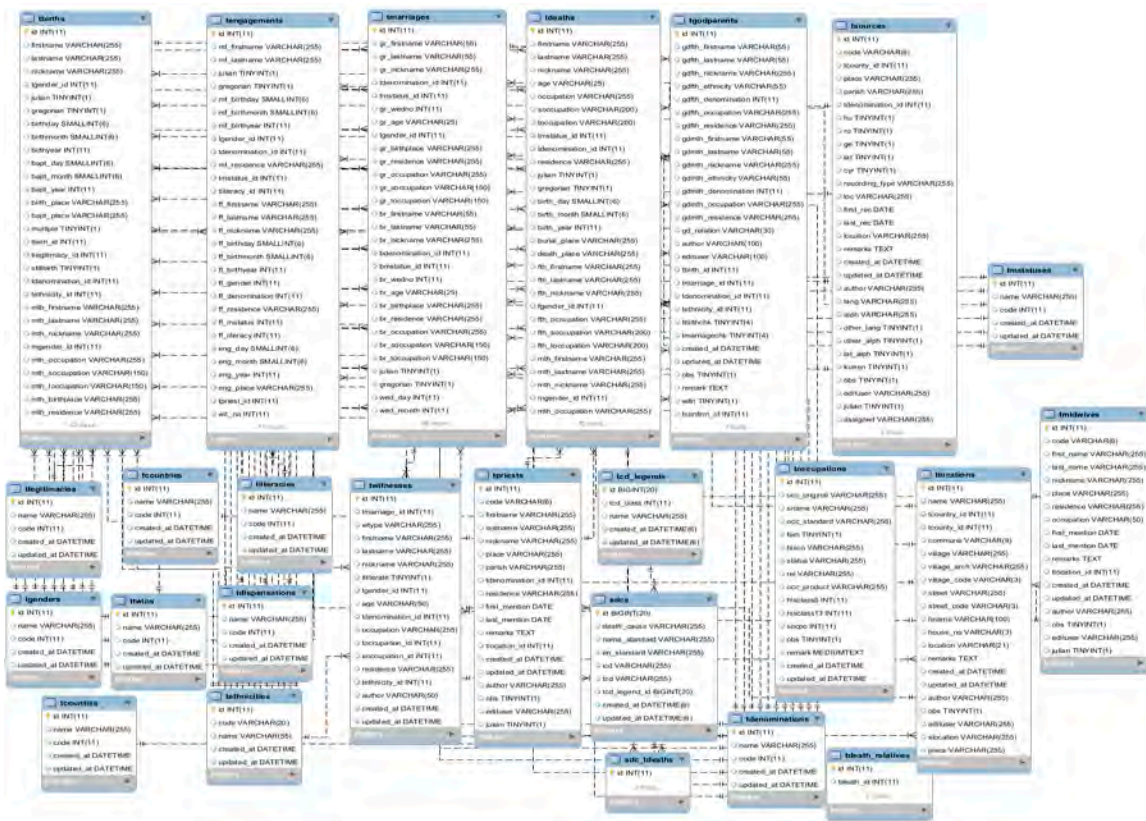


Figure 9 Model hpdt_v4



III

Semi-longitudinal data



HISTORICAL LIFE COURSE STUDIES
VOLUME 9 (2020), published 15-12-2020

The Richness of Italian Historical Demography

Marco Breschi
University of Sassari

Alessio Fornasin
University of Udine

Matteo Manfredini
University of Parma

ABSTRACT

In this paper, we present a new methodology for the reconstruction of individual life-histories based on information derived from the integration of different parish registers. This methodology makes it possible to associate the sequence and timing of demographic events not only with the structural features of the households in which they occurred, but also with more general historical context and the economic factors that shaped the lives of people and households. All these elements are then evaluated in a dynamic and temporal perspective, allowing the adoption of a longitudinal approach in the analysis of demographic processes for historical populations.

Keywords: Longitudinal databases, Life histories, Parish registers, Italy

DOI article: <https://doi.org/10.51964/hlcs9304>

© 2020, Breschi, Fornasin, Manfredini

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

Reconstruction of the demographic history of the Italian population can only be pursued at the cost of a great effort of research and commitment. The reason is not the limited availability of sources and historical data, as typical of other countries, but, paradoxically, the opposite. The problem was described by Karl Julius Beloch, the most important scholar of Italian historical demographic documentation and author of valuable essays on the history of the Italian population in the unsurpassed study *Bevölkerungsgeschichte Italiens* (1994)¹. In his short text advocating a new reconstruction of the Italian population (Beloch, 1887), Beloch notes that the sources are 'not a problem, indeed the richness and quantity of historical materials create major difficulties to those who want to undertake research in that field. Our archives are full and rich in documents and data providing precise knowledge of population(s), market prices, socioeconomic conditions, and other features since the 15th century' (Beloch, 1887, p. 48). He wrote this volume when he was 33 and had already launched his ambitious program of an extensive survey of demographic documents in all the national archives. The result in the following year (Beloch, 1888) was publication of his influential essay on the Italian population between the 16th and the 18th century.²

Since then, historical demographic research has made significant progress. In particular, basic demographic documentation, such as time series of population size as well as births and deaths, became the subject of important surveys and systematic analyses. Besides the work of Beloch, we must also note the extraordinary census of pre-unification (before 1861) archival sources launched by Corrado Gini, a project whose results were condensed in ten volumes published between 1933 and 1941.³ In this line of research, we also include scientific activity promoted by the Institute for Industrial Reconstruction (IRI). Their work supported a series of regional studies and, starting from the 1970s, an articulated series of seminars on demographic sources carried out by the Italian Committee for the study of Historical Demography (CISP) and by the newly founded Italian Society of Historical Demography (SIDES) after 1977.

There is no need to trace the development of the historical demographic research in Italy of the last 50 years here.⁴ However, despite undeniable advances and improvements, Beloch's remark that the richness of historical documentation has hindered the development of research on the history of the Italian population is still valid; the treasure trove of historical sources remains largely unexplored.

Because of the richness, complexity, and variety of demographic sources in Italian archives, we decided to focus on a set of parish registers existing almost everywhere in Italy. The systematic administration of demographic data by way of a population register and civil vital registration at the municipal level began after 1861, the year of Italian unification. Before that date, most of the old Italian states and kingdoms preferred to rely on the centuries-long practice of priests and parsons recording registers of baptisms, marriages, and burials as well as *Status Animarum*, a sort of annual census carried out on Easter. These four sources, which were systematically recorded since the 16th century, are available for most of the over twenty thousand parishes spread across Italy.⁵ The potential of such religious sources is well known. The most famous technique for reconstructing past populations, the family reconstitution method developed by Henry in the 1950s (Fleury & Henry, 1956), was based on these registers. It was used in hundreds of case-studies to describe both local and national demographic systems, such as the second volume of the history of the English population (Wrigley, Davies, Oeppen, & Schofield, 1997). The large majority of those studies relied exclusively on the three vital registers (baptisms, marriages, and burials), while very few used *Status Animarum*. This type of register is not common everywhere in

1 The work was published posthumously in three volumes between 1937 and 1961. It was translated into Italian and collected in a single book with an introduction by Lorenzo Del Panta and Eugenio Sonnino.

2 This essay represented a landmark for a long time: it was reprinted in 1959 in the first volume of the *History of the Italian Economy* edited by Carlo Maria Cipolla. In the next fifteen years, Cipolla (1965) and Bellettini (1973) published two important essays on the history of the Italian population using information reported by Beloch in the third volume of *Bevölkerungsgeschichte Italiens*.

3 Archival research directed by Gini was collected in the book series *Archival Sources for the Study of Population Issues* until 1848.

4 For a more extensive description of the development of the historical demographic research in Italy see the contribution by Lucia Pozzi and Eugenio Sonnino (2012).

5 The total number of parishes in Italy, according to the Italian census of 1881 and to the state boundaries of the time, was 20,465, which raised to 24,615 in 1951.

Europe, and it is not found in some areas of Italy. On the other hand, there are many Italian parishes where it is possible to find multiannual series of Status Animarum. The Italian Committee for the Study of Historical Demography emphasized the potential of studies combining these four parish registers in the 1970s (CISP, 1974). Unfortunately, few studies followed these recommendations.

Our paper aims to highlight the full potential and the advantages offered by the combined use of individual-level data from all four parish registers, and specifically:

- integrating the study of local communities;
- extending the analysis to larger areas;
- reconstructing long-term changes from the sixteenth century to the second post-war period (1961);
- integrating the study of individual and family life-histories;
- studying the mobility of individuals and families indirectly;
- incorporating further sources of economic, social, and/or cultural information.

In other words, through the combined use of the four types of parish register, demography in Italy can bring the lives of poor and humble people into history. Those people left small traces of their existence: a baptism, sometimes a marriage, some others a simple list of family members. By handling these fragments with care, the historical demographer can bring them back to life by reconstructing their experiences.

This essay takes up this challenge. The first part of the paper describes the methodology adopted to link information from the four parish registers and to reconstruct the socio-demographic system of a mid-19th-century sharecropping village, Casalguidi. Then, the analysis expands to reconstruction of the economic context by adding historical sources of economic conditions. The final section outlines further developments of the project, which will mainly expand the time frame.

2 THE STUDIED POPULATION

In the period 1819–1855, the community of Casalguidi belonged to the Grand Duchy of Tuscany, located a few kilometers from the cities of Florence and Pistoia. Despite its rural nature, Casalguidi could be viewed as a small town with an average population of 2,500 people, rather than a simple peasant village (Breschi, Derosas, & Manfredini, 2000). The local rural economy was largely dominated by sharecropping, the most typical form of land tenure of mid-19th century Tuscany. The features and conditions of this peculiar system of land tenure had a profound impact on members of sharecropping households, including their demographic behaviors. Compared to other socioeconomic categories, sharecroppers had higher fertility (Manfredini & Breschi, 2008), lower infant mortality (Breschi, Manfredini, & Pozzi, 2004; Manfredini & Pozzi, 2004), more complex household structures due to residence in joint families (Doveri, 2000; Poni, 1982), and a different level of nuptiality (Della Pina, 1990; Rettaroli, 1993). Depending on the length of sharecropping contracts, migration and mobility could be more intense than in a rural context dominated by smallholders (Breschi & Manfredini, 2002). The high levels of mobility of individuals and households living in Casalguidi were also sustained by the presence of day laborers and seasonal workers. Since farm laborers had very short-term contracts and did not live on a farm, their family system was largely based on simple and small nuclear households, in contrast to the households of sharecroppers (Barbagli, 1990).

Overall, the demographic system of Casalguidi was typical of *ancien régime* populations, characterized by high levels of mortality (life expectancy at birth was around 35 years on average) and high levels of fertility (the total marital fertility rate was about 8 children per married woman between 20 and 49 years). Nuptiality responded to a variety of socio-economic and cultural factors, and stricter control over marriage within sharecropping households produced higher levels of permanent celibacy (Derosas, Breschi, Fornasin, Manfredini, & Munro, 2014) This behavior was the consequence of the rigid norms regulating household organization and behavior imposed by sharecropping contracts and enforced by the strongly patriarchal structure of sharecropping households (Della Pina, 1990; Doveri, 2000; Giorgetti, 1974).

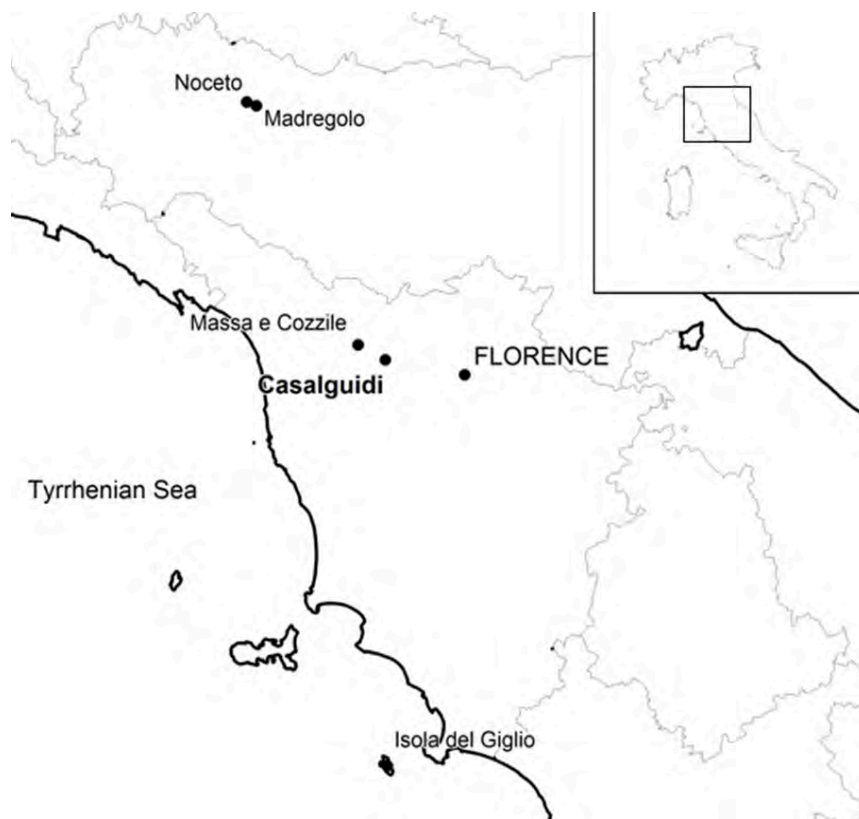
3 THE SOURCES

As already mentioned, the sources used to reconstruct the life histories of the inhabitants of Casalguidi were primarily of religious origin with the exceptions described below.

The parish registers of baptism, marriage, and burial became widespread across Italy at the end of the 16th century following the Council of Trent of 1545–1563, which made them compulsory for every person. Status Animarum were compulsory as well, but that duty was less respected by priests. Status Animarum were often of low quality, sometimes a simple list of household heads without any further information, sometimes a simple count of family groups and number of family members. Thus, Status Animarum are not common everywhere in Italy, and even where they are found, they often lack the necessary continuity over time. For this reason studies integrating parish registers and Status Animarum are rare in Italy. We can point to work by Carla Ge Rondi (1988) on a parish of the city of Pavia during the 18th century and a few others.⁶ We have personally collected such sources for a few parishes, but those of Madregolo and Casalguidi are, to our knowledge, the only ones for which a complete and integrated dataset has been created so far (see Figure 1).⁷ However, continuity over time of the Status Animarum has been the pre-condition that guided us in the choice of Casalguidi as a privileged case-study for our linkage methodology.

The parish registers of baptism, marriage, and burial show continuity over time between 1819 and 1859. The information contained in such registers are typical of Catholic records: the date of the event, name and surname of the recorded person(s), and father's and mother's names. Occupation and age were infrequently recorded, although the latter was often annotated for individuals who died at very young ages (see Table 1 for a complete overview).

Figure 1 *Geographical location of Casalguidi and the other parishes mentioned in the text*



6 See also the first volume by CISP (1974) on the sources of historical demography in Italy.

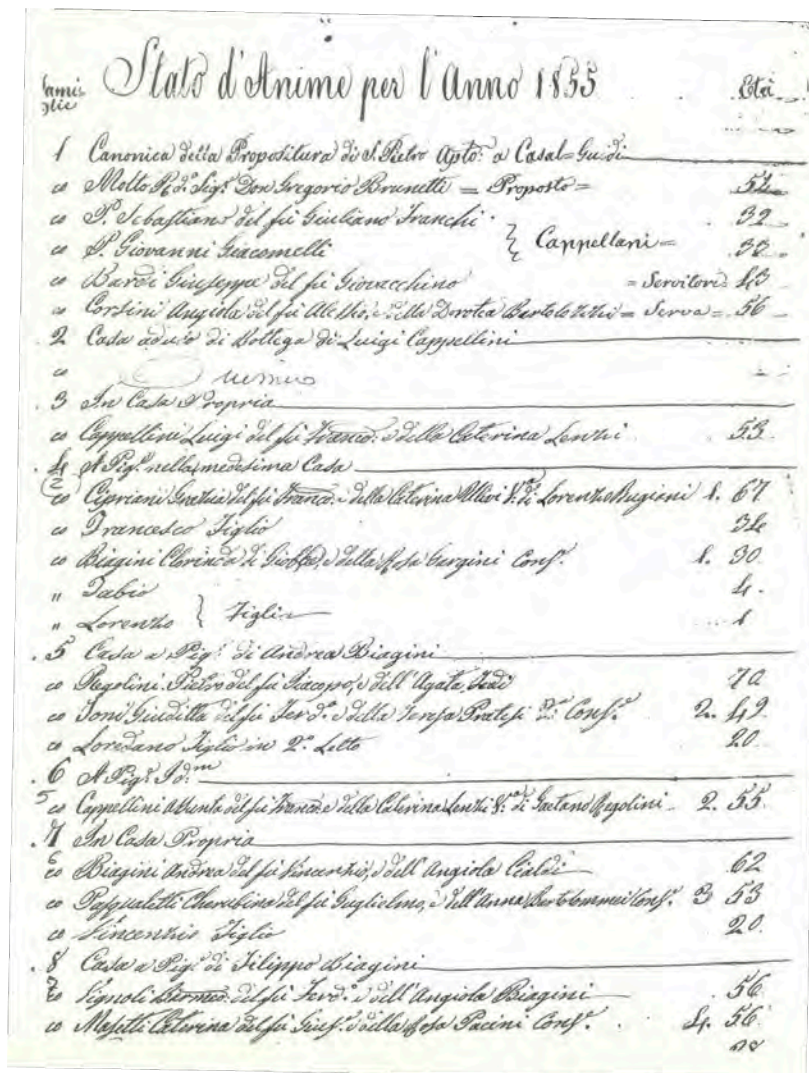
7 Other similar databases concern the parishes of Massa e Cozzile, Isola del Giglio, e Noceto.

Table 1 Available information by parish register

Information	Baptism	Burial	Marriage ^b	Status Animarum
Date of event	Always	Always	Always	Only Year
Place of event	Always	Always	Always	Always
Name	Often	Always	Always	Always
Surname	Always	Always	Always	Always
Age	Never ^a	Often	Often	Always
Father's name	Always	Always	Always	Often
Mother's name	Always	Often	Infrequent	Often ^c
Civil status	-	Often	Always	Always
Spouse's name	-	Infrequent	-	Often ^c
Relationship with household head	-	-	-	Always
Household head's profession	Often	Infrequent	Often	Often
House property	-	-	-	Always

^a Mother's age at birth; ^b Both spouses; ^c Inferred from the indication of relationships among members

Figure 2 Front page of Status Animarum, Casalguidi 1855



Between 1819 and 1859 only the Status Animarum of 1841 is missing, which can be replaced with information from the census of the Grand Duchy in that year. The Status Animarum (see Figure 2) provides a reliable and complete picture of the organization and structure of the households living in the parish. Who lived in the various households of the village? What were the relationships among family members? What were their demographic characteristics? Status Animarum answer all of these questions. If analyzed over time, in a diachronic perspective, they also provide researchers more detail about the mechanisms of household modification (fusion, fission, etc.), even tracing the movements of individuals between households. In this way, the study of the family shifts from a static view, simply looking at household composition in a given period, to a dynamic approach focused on the structural changes of households in relation to both internal and external events and/or stressors.

Status Animarum are even believed to provide a more reliable picture of the real population compared to the contemporary pre-unification population registers. Indeed, the list of the household members was recorded (updated) year after year, including people temporarily present. In contrast, the population registers recorded before 1861 suffered from delays in updating information and insufficient surveillance of temporary migrants.

In conclusion, the usual parish registers of baptism, marriage, and burial provide basic individual information (except for mother's age at birth) for the three essential demographic events. The Status Animarum allows us to recover information on the composition and structure of the household as well as on socioeconomic conditions (occupation of the household head and other members; property of the house; etc.). Moreover, the continuous updating of information makes it possible to analyze patterns of social mobility as well as household modification.

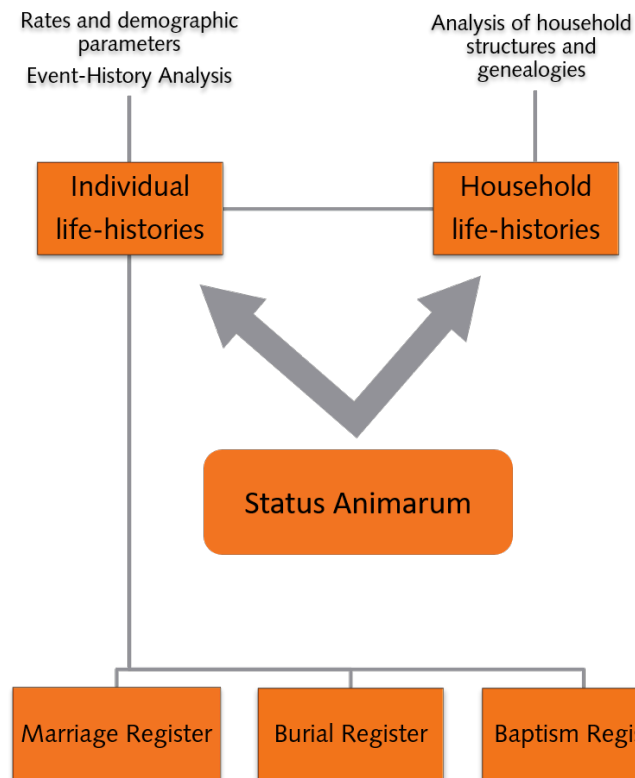
4 LINKAGE METHODOLOGY

From what has been said above, it is clear that Status Animarum and the vital parish registers complement each other.

The flow-chart of the linkage process is provided in Figure 3. The starting point is the nominative linkage between successive Status Animarum, which allows tracing the life history of each parishioner in Casalguidi through a unique individual identification code. This reconstruction includes the (changing) characteristics of the households in which he/she lived year by year. Nominative linkage identifies marriages between local men and non-local women, which were not recorded in the marriage register of the groom's parish due to the custom of celebrating the wedding in the bride's parish. Those unions can be identified in Status Animarum by the change in the groom's marital status between year t and year $t+1$.

The second step consists of linking the demographic events in the parish registers of baptism, marriage, and burial to each individual found in the Status Animarum. At this point, the reconstruction of the life history is completed, and it is now possible to frame each demographic event within the family and socioeconomic context in which it occurred.

From an operational point of view, the linkage strategy involved two stages. In the first phase we applied an automated linkage process in which records from the various sources were linked when the name, surname, paternity, and maternity matched exactly. A further check of internal coherence was then carried out on the year of birth (or age). After these links were removed from the original list of records, the remaining records underwent a semi-automatic process of nominative linkage, where we decided which records had to be linked from a sub-set selected by relaxing the above matching criteria. Obviously, the linkage process was not that easy and various problems occurred during the second phase. For example, the earliest records of deaths did not report the maiden name of married women but only the husband's surname; in the death and marriage acts, some individuals were indicated with their nickname, shortened name, or with a first name not reported in the baptism act.

Figure 3 *Linkage process for the reconstruction of life-histories*

Despite these problems, the benefits of this methodology are numerous and substantial, involving both qualitative and quantitative aspects, such as:

- interpreting individual behaviors in the light of household characteristics and vice versa;
- carrying out longitudinal analyses thanks to the reconstruction of individual and family life-histories;
- calculating rates and other demographic indices for cross-sectional analyses based on the population at risk for each specific event;
- analyzing household structure, its evolution, and its impact on demographic behaviors;
- studies of spatial and social mobility, both at the individual and family level.

The linkage methodology described above is also useful to check the internal consistency of data and to replace missing information in both sources. First, Status Animarum provide the ages of each individual for estimating the year of birth of people born outside the parish or in a period not covered by baptism registers. The more data on ages we found in Status Animarum, the more precise the estimation. Secondly, the annual pace of this sort of religious census allows us to determine with yearly precision the entry and/or the exit from observation (i.e. from Status Animarum) of each parishioner. The nominative linkage with baptism, marriage, and burial registers offers the demographic reasons for such movements. If an individual is no longer recorded on Status Animarum but is recorded on a burial or a marriage act, the exit from observation can be attributed to the individual's death or his/her marriage. If no event occurred, emigration is the most likely reason for the disappearance. Likewise, if a person entering Status Animarum is associated with a baptism, then their appearance in the record is due to birth and not to immigration. In addition, it is also possible to retrieve the age of individuals for all those events.

In conclusion, the integration of Status Animarum with parish registers allows reconstructing and/or improving estimates of the following information:

- Year of birth of each parishioner;
- Age at death for the burial acts not presenting such a piece of information;
- Individuation and analysis of the exogamous marriages celebrated elsewhere, in particular in the bride's parish. These unions remain completely unknown in classic family reconstitution (Manfredini, 2003), and strongly bias the analysis of nuptiality;

- Mother's age at birth for all births that occurred in the parish;
- Demographic characteristics of migrants and movers, which is an enormous step forward in historical demography, where the simple assessment of migrant stock is often impossible.

5 RECONSTRUCTING LIFE HISTORIES

The procedure described above allowed us to retrieve information on 8,015 individuals who resided in the parish of Casalguidi for at least one year. These people contributed for 96,581 person-years, an average of 12.1 years.

In order to reconstruct the life histories of those people, we first determined the year of occurrence of each demographic event. As shown in the tables below, the integration with Status Animarum made it possible to retrieve much more information. In particular, the year of birth was reconstructed from Status Animarum for about 62% of the parishioners. The contribution of Status Animarum to the year of death is more limited, around 7% of total deaths. When it comes to the year of marriage, the situation is more complex. As already mentioned, the custom in Italy was to celebrate the wedding in the bride's parish and to settle in the groom's parish. Consequently, the marriage year of males marrying outside the town could only be established from their change in marital status as noted in the Status Animarum (Manfredini, 2003). Table 2 shows that information retrieved from Status Animarum determined the year of marriage for about 29% of total unions.

Table 2 *Data retrieval by piece of information and source, 1819–59*

	Parish Register		Status Animarum		Not retrieved		Total	
	N	%	N	%	N	%	N	%
Year of birth	2,996	37.4	4,971	62.0	48	0.6	8,015	100
Year of death	2,171	92.9	166	7.1	0	0.0	2,337	100
Year of marriage ^a	580	70.9	238	29.1	0	0.0	818	100

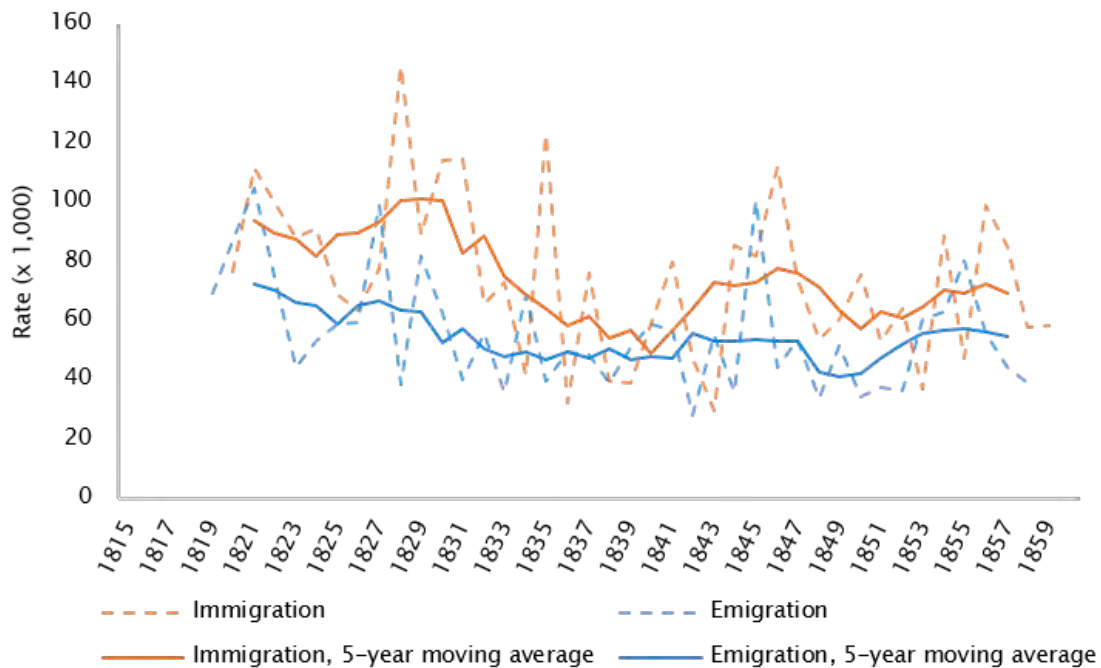
^a For couples

Infants who died within their first year of life may not have been recorded in the Status Animarum and they were therefore not included in table 2. In Casalguidi we found 490 infants in the baptism and burial records who do not appear in the Status Animarum.

After every event is assigned to a year, we can combine them into a life history, the sequence of demographic events characterizing the life of an individual. Table 3 clearly shows that year of birth is the only event observed for a large part of the population (42%).

Table 3 *Life-histories: reconstructed sequences of events*

Linked information	Casalguidi	
	N	%
Birth, marriage & death	627	7.4
Birth & marriage	2,057	24.2
Birth & death	2,200	25.9
Only birth	3,574	42.0
Residual combinations	4	0.04
No information	43	0.5
Total	8,505	100.0

Figure 4 *Migration rates, Casalguidi, 1819–59*

The sub-group only observed at birth is very large, because of a high migration rate, which was typical of rural Italian communities where most people did not own land (see Figure 4).

Another benefit of combining parish registers and Status Animarum is the study of migration and mobility, which is little studied in historical demography, especially from an individual perspective (Corsini, 1993; Levi, Fasano, & Della Pina, 1990). Migration, defined as the exit from the parish territory, is measured by the disappearance of a person between two successive Status Animarum without a record in the burial register. This methodology ensures reliable estimates of movements to and from the parish, and it can be used for movements between households in the same parish.⁸

Complete sequences from birth to marriage to death account for only 7.4% of life histories. The short time-span investigated (40 years) and low life expectancy at birth (around 35 years) make coverage of an entire life span very unlikely.

6 TWO MORE STEPS: LINKAGE WITH THE TAX REGISTER AND GRAIN PRICES

Recently, the analysis of economic differentials has been one of the most fertile research fields in historical demography. Analysis of the relationship between living standards and demographic events opened new frontiers in the study of demographic processes and their evolution (Allen, Bengtsson, & Dribe, 2005; Bengtsson, Campbell, & Lee, 2004). However, it is not easy to construct or calculate reliable indicators of the living standards and economic conditions of past populations. Occupation is usually the only source of economic information in parish registers, but it is affected by potential distortions (Manfredini, & Breschi, 2008). First, occupation is not recorded in a systematic way, and high-status occupations are sometimes privileged. Second, the socioeconomic categories adopted in the classification of occupations are often ambiguous. The definition of 'farmer' is a case in point, as it includes a multiplicity of different farm workers regardless of landowning and type of contract. Moreover, the occupation does not always reflect the real living standard of a person or a family:

⁸ The integration between parish registers and Status Animarum can also be used to analyse return migration. Although not specifically addressed so far, the linkage methodology makes it possible to identify individuals coming back to Casalguidi after a previous departure.

some sharecroppers experienced better economic conditions than many smallholders. Communities with strong seasonal migrations (Fornasin, 2005), in which people frequently had two or more jobs according to the season, are even more difficult to classify. Finally, information on occupations is not updated with sufficient frequency when derived only from parish registers.

To evaluate with more precision the standard of living of individuals and families we integrated the newly reconstructed life histories with information drawn from the tax register. This register was updated annually, and records for Casalguidi are continuous over time. The tax register reports the name and surname of the household head, his occupation, place of residence, household size, and the amount each household had to pay. Each household was assigned to a tax class based on economic and household indicators, such as total family income, number of household members, ownership of land and/or houses, etc. The number of tax classes varied over time.

Operationally, we decided to classify households into four tax categories: households with high, medium, and low tax as well as households exempted from payment for manifest poverty, such as unmarried women with children.⁹ The tax register was linked to the Status Animarum by nominative linkage through the head of the household to create a precise and continuous description of the socioeconomic status and living standard of each household. Table 4 shows the distribution of tax categories by household head's occupation. The relationship between occupation and standard of living is not as strong as expected, in particular, the assumption that day laborers lived in worse economic conditions than sharecroppers does not seem to hold.

Table 4 *Households by tax class and household head's occupation (%)*

Occupation	High	Medium	Low	Exempt	Total	
					%	N
Day laborers	0.9	11.6	58.6	28.9	100.0	4,723
Sharecroppers & smallholders	1.3	10.7	59.3	28.8	100.0	8,621
Non-rural activities	1.9	13.6	56.9	27.6	100.0	2,432
Nobles & the bourgeoisie	47.4	23.9	13.5	15.1	100.0	384
No profession	0.3	4.2	13.0	82.4	100.0	2,393
Total %	2.1	10.7	51.9	35.3	100.0	18,557
Total N	386	1,993	9,625	6,549	-	18,557

Finally, the last step was to integrate information on family socioeconomic status with an indicator describing the general economic context. Therefore, we collected information on annual and monthly prices of grain on the Florence market. Widely used in economic history, price series have acquired a key role also in historical demography (Bengtsson & Reher, 1998; Breschi, Fornasin, & Gonano, 2005; Breschi, Fornasin, & Manfredini, 2011). A rise in grain prices is usually an indicator of short-term crisis. Crises may have been purely economic or epidemiological in origin, because epidemics usually had important effects on markets. Analysis of the combined effects of general economic conditions and household socioeconomic status on demographic events has been one of the central issues of the Eurasian Project on Individuals and Households (Bengtsson, Campbell, & Lee, 2004). The results of that project highlighted differences among economic strata in reactions and resilience in times of economic stress.

9 The decision about the number of tax classes to analyze depends obviously on the specific research goals.

7 FUTURE PERSPECTIVES

In addition to reconstructions of other Italian populations and communities using the life history methodology described here, our team is working to expand the time-frame for the population of Casalguidi. We have already started to collect and computerize parish registers from the second half of the seventeenth century to 1819 and from 1859 to the first half of the 20th century. This project also involves exploitation of data from post-unification national censuses, in partial compensation for the disappearance of Status Animarum in the 20th century.

For Casalguidi, we are integrating additional sources with the reconstructed life histories, both at the individual and context level. Among the former, we can mention registers of physical examinations for military service, available for all 20-year-old men born from 1841 onwards. This source offers a way to analyze the health of the male population (Fornasin, Breschi, & Manfredini, 2017). In particular, we can use stature to infer the evolution of living standards as well as the role of physical characteristics on the risk of dying or getting married (Manfredini, Breschi, Fornasin, & Seghieri, 2013).

The digitization of cadastral registers will enable us to georeference both individual and household demographic data at the level of single houses. This could be very useful in studies of epidemics and marriage patterns (Corsini & Fornasin, 2017; Fornasin, Breschi, & Manfredini, 2016).

At the contextual level, we are enhancing our database with daily meteorological data, available continuously from the second half of the 18th century in the registers of the Ximenian observatory of Florence. These data will be valuable for evaluating the impact of weather conditions on the mortality of the most fragile individuals (infants and the elderly) (Scalone & Samoggia, 2018).

Future work will exploit the temporal continuity of the Status Animarum to capture dynamic processes within and between families and households. Examples are:

- spatial and economic aspects of relationships and ties among households;
- links between social and cultural changes in the institution of the family and the status of never-married or widowed individuals;
- demographic and economic inequalities within the household.

8 CONCLUSIONS

In this paper, we have stressed a new methodology for the reconstruction of individual life histories based on information derived from the integration of different parish registers. These methods make it possible to associate the sequence and timing of demographic events with the structural features of households and with broader contextual and economic factors shaping the lives individuals and households. These elements provide dynamic and temporal perspectives, allowing the adoption of a longitudinal approach in the analysis of demographic processes in historical populations.

In conclusion, the longitudinal database of the population of Casalguidi stands out for the richness and quality of its demographic, economic, and social information. The methods described here produce historical data supporting analyses comparable to those performed on contemporary populations.

REFERENCES

- Allen, R. C., Bengtsson, T., & Dribe, M. (2005). *Living standards in the past: New perspectives on well-being in Asia and Europe*. Oxford: Oxford University Press. doi: [10.1093/0199280681.001.0001](https://doi.org/10.1093/0199280681.001.0001)
- Barbagli, M. (1990). Sistemi di formazione della famiglia in Italia. In SIDES, *Popolazione, società e ambiente. Temi di demografia storica italiana (secoli XVII-XIX)* (pp. 3–44). Bologna: Clueb.
- Bellettini, A. (1973). La popolazione italiana dall'inizio dell'era volgare ai nostri giorni. Valutazioni e tendenze. In E. Castelnovo, G. Pestelli, R. Tedeschi & R. Leydi (Eds.), *Storia d'Italia* (Vol. 5, pp. 489–532). Turin: Einaudi.

- Beloch, G. (1887). Una nuova storia della popolazione d'Italia. *Nuova Antologia*, 22(17), 48–61.
- Beloch, G. (1888). La popolazione d'Italia nei secoli XVI, XVII e XVIII. *Bulletin de l'Institut International de Statistique*, 3(1), 1–42.
- Beloch, K. J. (1994). *Storia della popolazione d'Italia*. Firenze: Le Lettere.
- Bengtsson, T., Campbell, C., & Lee, J. Z. (2004). *Life under pressure: Mortality and living standards in Europe and Asia, 1700–1900*. Cambridge, MA: MIT Press.
- Bengtsson, T., & Reher, D. (1998). Short and medium term relations between population and economy. In C.-E. Núñez (Ed.), *Proceedings Twelfth International Economic History Congress: Debates and controversies in economic history* (pp. 99–115). Madrid: Fundación Ramon Areces.
- Breschi, M., Derosas, R., & Manfredini, M. (2000). Infant mortality in nineteenth-century Italy: Interactions between ecology and society. In T. Bengtsson & O. Saito (Eds.), *Population and economy. From hunger to modern economic growth* (pp. 457–490). Oxford: Oxford University Press.
- Breschi, M., Fornasin, A., & Gonano, G. (2005). Short-term demographic changes in relation to economic fluctuations: The case of Tuscany during the pre-transition period. In R. C. Allen, T. Bengtsson & M. Dribe (Eds.), *Living standards in the past. New perspectives on well-being in Asia and Europe* (pp. 319–340). Oxford: Oxford University Press. doi: [10.1093/0199280681.001.0001](https://doi.org/10.1093/0199280681.001.0001)
- Breschi, M., Fornasin, A., & Manfredini, M. (2011). Demographic responses to short-term stress in a 19th century Tuscan population: The case of household out-migration. *Demographic Research*, 25(16), 491–512. doi: [10.4054/DemRes.2011.25.15](https://doi.org/10.4054/DemRes.2011.25.15)
- Breschi, M., & Manfredini, M. (2002). Individual and family mobility. First results from an analysis on two Italian rural villages. In D. Barjot & O. Faron (Eds.), *Migrations, cycle de vie & marché du travail* (pp. 43–64). Paris: Société de Démographie Historique.
- Breschi, M., Manfredini, M., & Pozzi, L. (2004). Mortality in the first years of life: Socio-economic determinants in an Italian nineteenth-century population. In M. Breschi & L. Pozzi (Eds.), *The determinants of infant and childhood mortality in Europe during the last two centuries* (pp. 123–137). Forum: Udine.
- Cipolla, C. M. (1965). Four centuries of Italian demographic development. In D. V. Glass & D. E. C. Eversley (Eds.), *Population in history. Essays in historical demography* (Vol. 2: Europe and United States, pp. 570–587). London: Edward Arnold Publisher.
- Comitato italiano per lo studio della demografia storica (CISP) (1974). *Le fonti della demografia storica in Italia. Atti del seminario di demografia storica 1971–1972* (Vol. 1, Part 1). Rome: CISP.
- Corsini, C. A. (1993). Le migrazioni interne e a media distanza in Italia: 1500–1900. *Bollettino di Demografia Storica*, 19(1), 9–27. Retrieved from <http://www.demostorica.it/index.php/bollettino-demografia-storica.html>
- Corsini, C. A., & Fornasin, A. (2017). I matrimoni e la distanza matrimoniale nel Granducato di Toscana (1840–42). *Popolazione e Storia*, 2, 9–25. Retrieved from <https://popolazioneestoria.it/issue/view/59/showToc>
- Della Pina, M. (1990). Famiglia mezzadrile e celibato: Le campagne di Prato nei secoli XVII e XVIII. In SIDES, *Popolazione, società e ambiente. Temi di demografia storica italiana (secoli XVII–XIX)* (pp. 125–139). Bologna: Clueb.
- Derosas, R., Breschi, M., Fornasin, A., Manfredini, M., & Munro, C. (2014). Between constraints and coercion. Marriage and social reproduction in northern and central Italy in the eighteenth and nineteenth centuries. In C. Lundh & S. Kurosu (Eds.), *Similarity in difference: Marriage in Europe and Asia, 1700–1900*. Cambridge, MA: MIT Press. doi: [10.7551/mitpress/9780262027946.003.0009](https://doi.org/10.7551/mitpress/9780262027946.003.0009)
- Doveri, A. (2000). Land, fertility, and family: A selected review of the literature in historical demography. *Genus*, 56(3/4), 19–59. Available from <https://www.jstor.org/stable/29788654>
- Fleury, M., & Henry, L. (1956). *Des registres paroissiaux à l'histoire de la population; Manuel de dépouillement et d'exploitation de l'état civil ancien*. Paris: INED.
- Fornasin, A. (2005). Escaping the crisis. Friulan mountains 18th–19th century. *International Journal of Anthropology*, 20, 299–306. doi: [10.1007/BF02443065](https://doi.org/10.1007/BF02443065)
- Fornasin, A., Breschi, M., & Manfredini, M. (2017). Le mappe di salute di una popolazione storica prime indagini sul Friuli (XIX secolo). *Acta medico-historica Adriatica*, 15(1), 31–50. doi: [10.31952/amha.15.1.2](https://doi.org/10.31952/amha.15.1.2)
- Fornasin, A., Breschi, M., & Manfredini, M. (2016). Environment, housing, and infant mortality: Udine, 1807–1815. In D. Ramiro Fariñas & M. Oris (Eds.), *New approaches to death in cities during the health transition* (pp. 43–54). Springer. doi: [10.1007/978-3-319-43002-7_3](https://doi.org/10.1007/978-3-319-43002-7_3)
- Ge Rondi, C. (1988). *L'analisi nominativa in demografia storica: metodi e problemi. Il caso di una parrocchia*. Milano: Giuffrè Editore.

- Giorgetti, G. (1974). *Contadini e proprietari nell'Italia moderna: Rapporti di produzione e contratti agrari dal secolo XVI ad oggi*. Turin: Einaudi.
- Levi G., Fasano, E., & Della Pina, M. (1990). Movimenti migratori in Italia nell'età moderna. *Bollettino di Demografia Storica*, 12(1), 19–34. Retrieved from <http://www.demostorica.it/index.php/bollettino-demografia-storica.html>
- Manfredini, M. (2003). The use of parish marriage registers in biodemographic studies: Two case studies from 19th-century Italy. *Human Biology*, 75(2), 255–264. doi: [10.1353/hub.2003.0034](https://doi.org/10.1353/hub.2003.0034)
- Manfredini M., & Breschi, M. (2008). Socioeconomic structure and differential fertility by wealth in a mid-nineteenth century Tuscan community. *Annales de Démographie Historique*, 1, 15–33. doi: [10.3917/adh.115.0015](https://doi.org/10.3917/adh.115.0015)
- Manfredini M., Breschi, M., Fornasin, A., & Seghieri, C. (2013). Height, socioeconomic status and marriage in Italy around 1900. *Economics & Human Biology*, 11(4), 465–473. doi: [10.1016/j.ehb.2012.06.004](https://doi.org/10.1016/j.ehb.2012.06.004)
- Manfredini M., & Pozzi, L. (2004). Mortalità infantile e condizione socio-economica. Una riflessione sull'esperienza italiana fra '800 e '900. *Revista de Demografía Histórica*, 22(2), 127–156. Retrieved from <https://dialnet.unirioja.es/ejemplar/109160>
- Poni, C. (1982). La famiglia contadina e il podere in Emilia Romagna. In C. Poni, *Fossi e cavedagne benedicon le campagne: Studi di storia rurale* (pp. 283–356). Bologna: Il Mulino.
- Pozzi, L., & Sonnino, E. (2012). Demografia storica: Un secolo di ricerca in Italia. *Popolazione e storia*, 13(2), 129–182. Retrieved from <https://popolazioneestoria.it/issue/view/40/showToc>
- Rettaroli, R. (1993). Maritu a chi troa, moglie a chi tocca. Nuzialità e famiglia nell'Italia mezzadrile del primo Ottocento. In SIDES, *La popolazione delle campagne italiane in età moderna* (pp. 505–526). Bologna: Clueb.
- Scalone, F., & Samoggia, A. (2018). Neonatal mortality, cold weather, and socioeconomic status in two northern Italian rural parishes, 1820–1900. *Demographic Research*, 39(18), 525–560. doi: [10.4054/DemRes.2018.39.18](https://doi.org/10.4054/DemRes.2018.39.18)
- 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.

The Utah Population Database

A Model for Linking Medical and Genealogical Records for Population Health Research

Ken R. Smith	University of Utah
Alison Fraser	University of Utah
Diana Lane Reed	University of Utah
Jahn Barlow	University of Utah
Heidi A. Hanson	University of Utah
Jennifer West	University of Utah
Stacey Knight	Intermountain Healthcare, Salt Lake City, Utah
Navina Forsythe	Utah Department of Health
Geraldine P. Mineau	University of Utah

ABSTRACT

Improving our understanding of the socio-environmental and genetic bases of disease and health outcomes among individuals, families, and populations over time requires extensive longitudinal data on multiple attributes for entire communities, states or nations. This requirement can be difficult to achieve. In this paper we describe a successful example of a database that meets these needs. The Utah Population Database (UPDB) is a unique and powerful database rarely found in the world that has been addressing these data requirements for over 40 years. The UPDB at the University of Utah is one of the world's richest sources of in-depth information that supports research on genetics, epidemiology, demography, history, and public health. Genetic researchers have used UPDB to identify and study individuals and families that have higher than normal incidence of diseases or other traits, to analyze patterns of genetic inheritance, and to identify specific genetic mutations. Demographers and other social scientists are increasingly using the UPDB to study issues such as trends in fertility transitions and shifts in mortality patterns for both infants and adults. A central component of the UPDB is an extensive set of Utah family histories, in which family members are linked to demographic and medical information. The UPDB includes medical information about cancer, causes of death, and medical details associated with births. It also includes diagnostic records from statewide insurance claims data and healthcare facilities (hospital discharge, ambulatory surgery, emergency department encounters). UPDB is also linked to Medicare claims data, a federal health insurance program generally for persons age 65 or older. The UPDB provides access to information on more than 11 million individuals and supports nearly 400 research projects. We describe in detail the data components of the UPDB, how it can be accessed, issues related to its development, record linkage, governance and privacy protections, as well as plans for future developments.

Keywords: Historical demography, Demography of Utah, Record linking, Administrative records, Data privacy, Genetics

DOI article: <https://doi.org/10.51964/hlcs11681>

© 2022, Smith, Fraser, Reed, Barlow, Hanson, West, Knight, Forsythe, Mineau

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

A strategy for understanding human health and well-being is to collect and curate extensive data on large, well-defined populations and all of its members over time with the appropriate data and privacy safeguards. Several successful (some longstanding) examples of these databases exist, many of which have been fundamental contributors to key medical and social science discoveries. The Framingham Heart Study in the US and the National Survey of Health and Development in the UK are exemplary in this respect.

In this paper, we describe a unique resource, the Utah Population Database (UPDB), which offers exceptional and unique data and research opportunities for population scientists, demographers, epidemiologists, historians, geneticists, health services researchers, and behavioral scientists, among others, all of whom work on population health and medical research. A distinctive quality of the UPDB is that it is based on links at the individual-level of administrative and medical records derived from a range of sources spanning decades for some sources and centuries for others (Casey, Schwartz, Stewart, & Adler, 2016; Hurdle, Smith, & Mineau, 2013) with many sources updated up to the present. The individuals with linked records comprising life histories are in turn linked to their family members, a feature of UPDB that allows analysts to study families, shared and unshared environments, and genetic associations over many generations. Moreover, these linkages create up to 17 generations and the concept of family can be expanded such that many individuals are frequently connected to tens of thousands of relatives by blood or marriage. Members of these multi-generational pedigrees have extensive event and date information but also spatial attributes at varying levels of geographic detail. These latter data elements allow projects to link geo-coded data to the UPDB in order to investigate environmental exposures as well as factors related to propinquity such as the geographic distance separating relatives or travel time needed to access a hospital.

This article is structured around central features and characteristics of UPDB's history and its structure and management. The paper starts with the specific components that comprise it today and the historical circumstances that led to their inclusion in UPDB. We then describe our conceptual data model and how and why UPDB has been able to thrive and grow for so many decades, in short, due to consistent institutional commitments. This is followed by a section devoted to details about the record linking methods used to create UPDB. Given the sensitive nature of the data in UPDB and the need to maintain the highest level of data security, we describe the regulatory protections of the data and how research access to the data can be obtained. Confidentiality and privacy issues are discussed in the context of UPDB as well as how these matters relate to UPDB's relationship to the many agencies which provide data. The paper ends with final thoughts and directions for the future.

2 DATA

2.1 SOURCES

UPDB was established over 40 years ago and has been a premiere research resource that had the early vision to integrate genetics and the social sciences. Selected key dates representing important developments in the evolution of UPDB are shown in Figure 1. Its beginnings can be dated to the years 1973–1974 when several researchers at the University of Utah realized the research opportunities that could be gained by first obtaining extensive genealogy records and constructing a population-based resource that would link these genealogical data to high quality medical records in order to investigate the genetic basis of a number of important diseases. Central to the launch of UPDB was geneticist Mark Skolnick, who was recruited to the University of Utah to lead a computerization of family history records with links to medical records. He then created an initial consortium of two additional key scientists, cardiologist Roger Williams and demographer Lee L. Bean.

Figure 1 Utah Population Database (UPDB) - Selected events in the history of UPDB

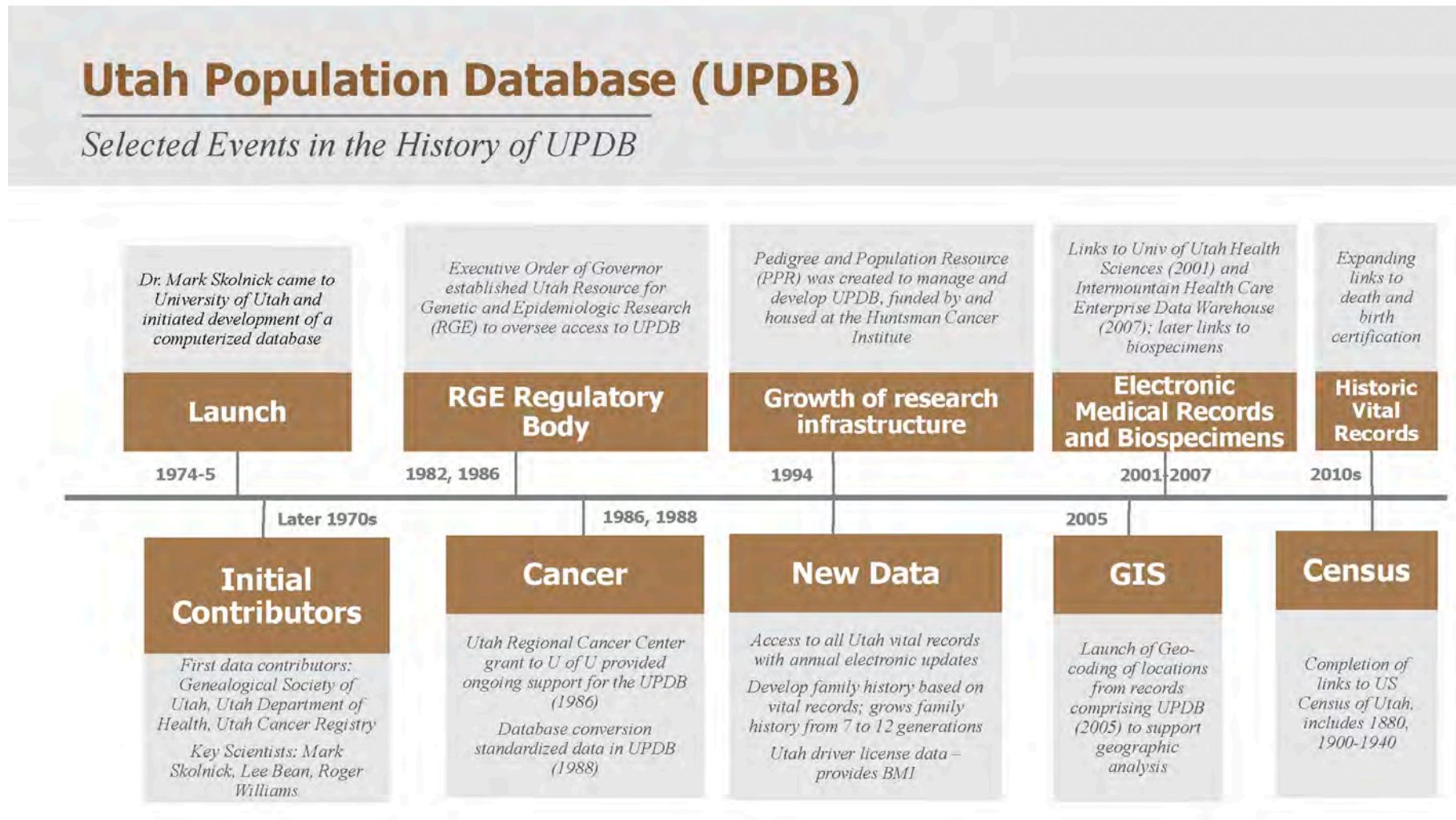


Table 1 *Population of Utah between 1850 and 2020*

Year	Population
1850*	11,380**
1860	40,273**
1870	86,786
1880	143,963
1890	210,779
1900#	276,749
1910	373,351
1920	449,396
1930	507,847
1940	550,310
1950	688,862
1960	890,627
1970	1,059,273
1980	1,461,037
1990	1,722,850
2000	2,233,169
2010	2,763,885
2020	3,249,879

* *Members of the Church of Jesus Christ of Latter-day Saints arrive in Utah July 24, 1847.*

** *Population of the Territory of Utah which included parts of present-day states of Colorado, Nevada, and Wyoming.*

Utah granted statehood January 4, 1896.

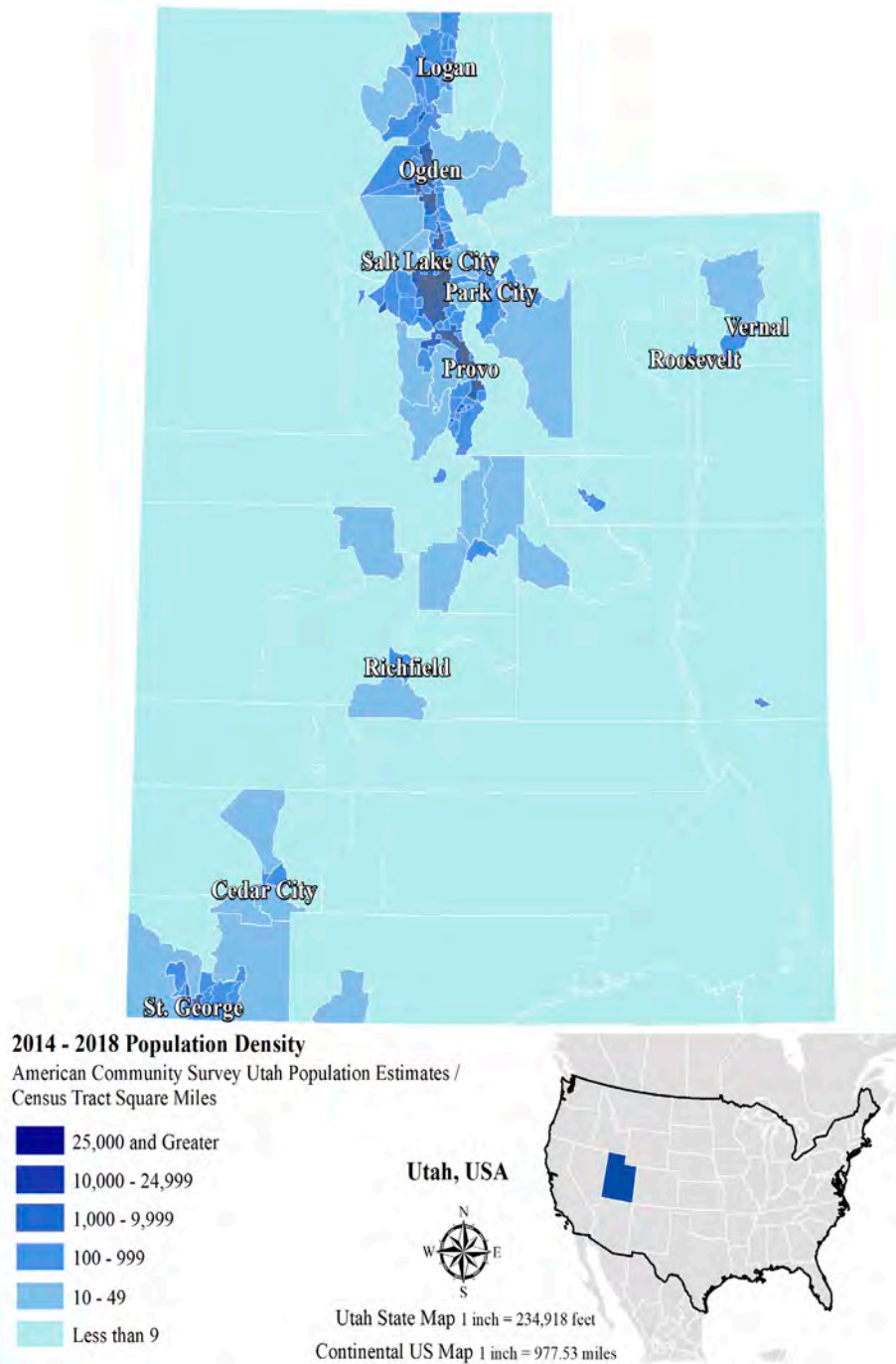
Sources: U.S. Census Bureau and Utah History Encyclopedia

UPDB is largely derived from records pertaining to events in Utah although connections to events outside the state are included when available, as described later. Utah is the 11th largest state in the US and has a median household income of \$71,414 (2018). Between 2020 and 2021 the population experienced a 1.8% increase. In Table 1 we show the growth of the population of Utah since its settlement in 1847, through the time Utah was admitted to statehood in 1896, and up to the present. Based on the 2020 US Census of Utah, Utah is comprised of 90.6% white (of which 77.8% are non-Hispanic), 1.5% African American/Black, 1.6% American Indian/Alaska Native, 2.7% Asian, 1.1% Native Hawaiian/Pacific Islander, and 2.6% two or more races; 14.4% are Hispanic or Latino. While Utah has 33.6 residents per square mile, it has the 8th highest percentage of people living in urban areas (2010 US Census) among the 50 states. Figure 2 illustrates how Utah has a low population density and high urbanization levels. Utah has a period life expectancy at birth of 79.9, which is above the US figure of 78.7 (2018).

The original set of genealogy records used when the UPDB was being developed comprised approximately 185,000 documents representing, on each form, three generations: a husband-wife pair, their four parents, and the couple's offspring and their respective spouses. These initial documents were selected to represent approximately 1.9 million individuals. Linking these across generations (e.g., a child in one group sheet is a parent on another) creates thousands of multi-generational pedigrees, providing astonishing insights regarding the population (Song & Campbell, 2017).

These early genealogical records comprise the original backbone of UPDB. The founding research team secured access to the Utah Cancer Registry (UCR, a Surveillance, Epidemiology and End Result (SEER) Registry) and Utah death certificates (from the Utah Department of Health) as the basis for medical outcomes to be linked to the genealogies at the individual level. Accordingly, many of the early studies focused on cancer based on these cancer records (Skolnick et al., 1981), as well as cardiovascular mortality (Williams et al., 1979) and demographic studies (Bean, May, & Skolnick, 1978; Skolnick et al., 1978) based on death and genealogy records (Skolnick, Bean, Dintelman, & Mineau, 1979).

Figure 2 Population density and map of Utah



The UPDB is a research resource that has been expanded extensively in its 40 years of existence. At this time, UPDB includes information on approximately 11 million individuals who have basic demographic information and is a data source for nearly 400 research projects. The time period within UPDB covers birth cohorts from the 1700s but are more extensive starting in the mid-1800s and run through the present. Using the UPDB Query tool (<https://uofuhealth.utah.edu/huntsman/utah-population-database/services/query.php>; requires registration) in December 2021, we show there are 304,104 individuals in UPDB born before 1847 (the year members of the Church of Jesus Christ of Latter-day Saints first arrived in Utah), 975,081 born between 1847–1899, 2,491,970 born between 1900–1949, 2,494,871 born between 1950–1974, 2,814,316 born between 1975–1999, and 1,734,962 born between 2000–2020, the latest update.

While the original development of UPDB was derived from three sources (genealogy, cancer record and death records), the UPDB now includes substantially more records and from diverse sources (see Table 2). These are fully described at <https://uofuhealth.utah.edu/huntsman/utah-population-database/>.

Table 2 *Records available in the Utah Population Database*

Record Type	Years Available	Notes	Records
Original Family History Records	1700's–1975	The original genealogical portion of UPDB holds Utah family histories organized into pedigrees based on Genealogical Society of Utah documents that hold demographic/kinship data.	1,917,111
UTAH VITAL RECORDS			
Birth Certificates	1915–1921, 1926–2020	Data on parents and children and their demographic medical information; volume of data varies by year.	3,162,090
Death Certificates	1904–2021	Causes of death are coded using International Classification of Diseases (ICD) revisions 6–10.	982,662
Marriage Certificates	1978–2010	Husband and wife name and age, marriage date, and county of marriage; (1988+): birth date and birth place, education, number of marriages, and type of marriage (civil/religious).	692,838
Divorce Records	1978–2010	Husband and wife name, marriage and divorce dates, and county where the divorce was issued; (1988+): birth dates and education of the husband and wife, number of marriages, number of children, and number of children under age 18.	298,928
Fetal Deaths	1978–2020	Stillbirths/fetal deaths of 20 weeks or greater gestation as calculated from the mother's last normal menses period to the date of delivery.	11,933
MEDICAL RECORDS			
Ambulatory Surgery Utah	1996–2020	Diagnosis and procedure codes and external injury E-codes.	12,342,203
Inpatient Hospital Claims Utah	1996–2020	Diagnosis and procedure codes and external injury E-codes.	6,696,825
Emergency Department	1996–2020 (older records forthcoming)	Diagnosis and procedure codes and external injury E-codes.	16,167,073
All Payer Claims Data	2013–2020	The APCD data captures medical and financial information for nearly all encounters involving 3rd party payers.	>200,000,000
Utah Cancer Registry	1966–2019	UCR is a statewide cancer registry that monitors cancer incidence & mortality. It participates in the NCI Surveillance, Epidemiology, and End Results (SEER) Program.	420,185
Birth Defect Network	1995–2018		23,910
ADDITIONAL RECORDS			
U.S. Census of Utah	1880,1900–1940	Individual-level records provide a range of data including SES, household composition, migration, literacy, and neighborhoods.	2,300,084
Social Security Death Index	Last updated 2011	Date and state of deaths regardless of their place of death.	581,373
Utah Driver License Division	Last updated 2021	DLD has residential data for all Utah drivers. DLD is also a good source for height and weight, or BMI.	4,175,080
Utah Voter Registration	Last updated 2020	Variables include residential information; updated during presidential election years.	2,251,922
TOTAL			>252,024,217

Externally Linked Records — Demographic records of external records that are linked to UPDB but substantive variables are held by data provider until investigators obtain IRB Approval			
Record Type	Years Available	Notes	Individuals
University of Utah Health Sciences	1992–Current	All University of Utah inpatient and out-patient clinics. Demographic records and medical and pharmacy information.	3,232,154
Intermountain Healthcare (IH)	1992–Current	All IH hospitals inpatient and out-patient clinics. Demographic records and medical and pharmacy information.	8,141,654
Centers for Medicare & Medicaid Services	1992–2012	Diagnostic, procedure and other risk factor data. 18 years of data linked to UPDB; new links are underway for more recent records.	700,000
Utah Department of Human Services (DHS)	1995–2020	DHS comprises 12 divisions including the child-serving Divisions of Child and Family Services (DCFS), Juvenile Justice Services (DJJS) and Services for People with Disabilities (DSPD). Data include records related care as supervision, wraparound services while in custody, therapeutic services, and in-home services.	616,894

Briefly, in addition to the original genealogy records, UPDB includes:

1. All electronically available Utah vital records (births, deaths, marriages, divorces and fetal deaths) from the Utah Department of Health from 1904 at the earliest onwards depending on the certificate.
2. Statewide health data from the Utah Department of Health including:
 - a. Ambulatory Surgery records which contain medical, financial and diagnostic information regarding visits occurring at designated surgical out-patient units;
 - b. Inpatient Hospital Discharge records which contain medical, financial and diagnostic data upon discharge from a hospital as an inpatient;
 - c. Emergency Department records which describe the medical and diagnostic information about the emergency visit;
 - d. All Payer Claims which hold data on medical, financial, diagnostic and pharmacy data that involve claims to a third-party health insurance provider;
 - e. Utah Cancer Registry data from the Utah Department of Health which hold statewide medical data on all incident cancer diagnoses except non-melanoma skin cancer;
 - f. Utah Birth Defect Network is a statewide, population-based surveillance system that identifies birth defects in children born in Utah since 1994; UPDB has data up to 2018.
3. Social Security Death Index records which provide place and date of death for persons who have ever been enrolled in the Social Security system.
4. Utah Voter Registration which provides information about whether still living in Utah.
5. Utah driver license records which contain data on spatial information on the place of residence as well as height and weight.
6. The 1880 and 1900–1940 Utah individual-level censuses.

UPDB is noteworthy with respect to linkages to other large federated medical data sets (i.e., links to UPDB but the information does not reside within UPDB). First, a “master subject index” or MSI has been created that links the UPDB with the demographic records from the Enterprise Data Warehouses (EDWs) of the two largest health providers in Utah: University of Utah Health Sciences and Intermountain Healthcare (DuVall, Fraser, Rowe, Thomas, & Mineau, 2012). These linkages are based on demographic information only and do not involve in any way medical, treatment or diagnostic information for the purposes of record linking. These two health care providers represent

inpatient and outpatient electronic medical information for approximately 85% of the state's medical encounters starting from the mid-1990s. These medical data are not held within the UPDB but are securely maintained by the enterprise data warehouses of these health care providers. Medical data are joined with the demographic and genealogical data in UPDB after the research project receives the necessary approvals from appropriate Institutional Review Boards (IRB) and the Utah Resource for Genetic and Epidemiologic Research (RGE), which oversees research access to the UPDB as described below.

A third and related medical data linked to the UPDB are those derived from Medicare claims, a federal health insurance program generally for persons age 65 or older. The Medicare data are available due to funding from National Institutes of Health (NIH) grants that were originally designed to facilitate the study of healthy aging and health expectancy among the Medicare-eligible (age 65 or older) population. These data relate to claims from 1992–2015 and more recent years are being added, if an individual had a claim at some point in Utah. These data are available to researchers beyond its original purposes using the UPDB but they must not only obtain IRB approval for their use but also approval from the federal Centers for Medicaid and Medicare Services (CMS).

Another data set linked to the UPDB stems from the Department of Human Services (DHS). The DHS data represent nearly all persons in Utah identified by the DHS including those using Aging and Adult Services, Child and Family Services, and Juvenile Justice Services. The linking also uses the “master subject index” methodology. Demographic data of included persons are provided for linking without any indication of the reason regarding their inclusion in the dataset. With DHS, RGE, and IRB approval, all requested information in each DHS purview is provided to the researcher.

2.2 RESEARCH OPPORTUNITIES

Linking these records within UPDB creates diverse types of datasets with unique research opportunities including:

1. Creation of reproductive histories

Using data from the Utah Department of Health that includes Utah birth certificates from 1915 to the present, we have extended the genealogical holdings of UPDB considerably. Information for the same mother and/or the same father on multiple birth certificates are linked, a technique similar to family reconstitution. This allows us to see that specific individuals share common parents and are therefore siblings. The children named on these birth certificates (the second generation) are then linked to the birth certificates of their children (the third generation, that is the grandchildren of the first generation). This provides an efficient and non-biased approach for representing the current Utah population as these families propagate the next generation. For the many families that remain in the state over their reproductive years, a complete history is possible. Moreover, this strategy creates broader genealogies connecting individuals more distantly related. Because birth certificates provide gestational age and birth weight as well as other features such as adverse obstetric events and birth complications, this strategy has provided a valuable source for analysis of preterm births, cesarean sections and preeclampsia in families and across generations (Hammad et al., 2020; Theilen et al., 2016, 2018). It is noteworthy that many of the genealogies derived from vital records also link into the legacy genealogies that are part of the UPDB. Note that this strategy is restricted to births visible on Utah birth certificates. Certainly, instances exist where a woman or a couple will bear children in Utah and others who were born elsewhere. Some data about past fertility patterns are represented on each birth certificate such as the number of previous pregnancies and live births (the availability of these data varies by birth year). This type of “retrospective” information from birth certificates is captured in UPDB but is not useful for constructing and expanding genealogies since the identities of these previous offspring born elsewhere are not known via birth certificates. Note that other sources of data, such as from the Genealogical Society of Utah or death certificates in UPDB that identify other offspring are used whenever possible.

2. Creation of residential exposures and histories.

Residential location information is derived from several sources in UPDB including Driver License Division (DLD) data, voter registrations, and vital records, while other records provide location information at a higher level of geographic aggregation such as the ZIP code. One use of DLD is to provide current residence status for individuals in UPDB. In this way a researcher is able to determine if an individual

is currently under observation, while residence information on death records verify if an individual was under observation until their death. This helps with generating population denominators. Every four years after major federal elections, Voter Registration records are obtained and linked to UPDB which give geographic information at a particular point in time. Additionally, DLD data hold information on height and weight from which we have derived the Body Mass Index (BMI) for each individual (Chernenko, Meeks, & Smith, 2019; Smith et al., 2008; Smith et al., 2011; Zick et al., 2009). Addresses from any of its sources used to derive residential histories within UPDB have been geo-coded when sufficient address information is available in the source records. This creates the opportunity for linking any geo-referenced data set (e.g., census block, air quality monitors) with individual-level data. These residential histories can capture important points in the life history of an individual from mother's residence at birth (own birth certificates), residence in childhood (birth certificates of latter born siblings), place of residence of offspring (children's birth/fetal death certificates), adult locations (census records, DLD, voter registration, health facilities data), and death (own or spouse: death certificates). Finally, residential histories described here refer to places within the state of Utah. Some records, including data from the Genealogical Society of Utah and the U.S. Census of Utah, contain information about locations for individuals outside the state of Utah. These may refer to places that precede or follow a period of time when an individual was living in Utah. In addition to that, to deal with potential selective migration into and out of Utah, UPDB staff have created date variables that mark the point in time when we first saw individuals and when we last saw them in the state of Utah, subject to the data availability within UPDB. For minors who do not yet vote or drive, mother's information is used. When possible, prior or subsequent specific locations are available in UPDB but even when they are unavailable analysts may use our entry and exit dates in order to adjust for possible selection bias. For analyses that span historic periods covered by U.S. Census records linked to UPDB, decennial sightings of locations are available whether or not they occur within Utah's boundaries.

3. Creation of Links with Individual-Level Census Records

The addition of the micro level census records from 1880 and 1900–1940 (and those to come) to UPDB now allows for several types of studies. First, it is now possible to observe mobility, both geographic and socioeconomic, and its causes and consequences. Seeing the population before the censuses and decades after the last one in 1940 enable investigators to see how personal fortunes (or penury) during these early years as reflected in the Census are associated with later life health and well-being. Second, given the manner in which census enumerators were assigned to districts to conduct the full count of the population, the data can be used to cluster individuals into neighborhoods. Accordingly, individuals identified in the census can be characterized by the quality of their 'neighborhoods' and how these spatial attributes may alter later life outcomes. These census records provide valuable independent information about family composition, co-residence, and genealogical data that may not be possible from other sources of data in the UPDB. Again, in terms of residential history, the censuses add value since they provide information about birthplace (important for the 19th century since Utah was greatly affected by international in-migrants) and in some cases (1910 Census) the year of entry to the US.

4. Creation of a Life Course Dataset to measure adversity and opportunity over time

With administrative data linked over many decades, the possibility of conducting life course analysis at the population level grows substantially. Since UPDB holds data from its earliest years in the 18th and 19th centuries up to the present, it is possible to see entire life spans within individuals and across generations. Apart from linking basic demographic and genealogical connections, UPDB annotates these records with information from vital records starting in the early 20th century, adds micro-level census information from 1880–1940, introduces cancer incidence information in the mid-1960s and then grows to include more medical data from the mid-1990s to the present. Family connections and geographic information exist throughout these years, though the spatial data vary in terms of their geographic resolution given the type of records available in a given period. UPDB has been the basis for the Demographic Child Adversity Exposure (DECADE) scale (Hollingshaus, 2015) which measures how challenges early in life may be associated with serious health outcomes, such as suicide, later in life. For contemporary years where data-rich birth certificates are available, other indicators of socioeconomic status include education (after 1968) and occupation (all years through 2008) of the parents are available in UPDB, along with marital status. This enables investigators to see children born into single-parent households or to find same-sex unions (for the latter, this has only been possible in recent years and is under development). These variables can be used to examine the effects of early life adversity on life courses for modern decades (Stroup et al., 2017).

5. Creation of Datasets with Links to External Datasets

UPDB also has the capacity to link its data to ongoing projects that have arisen independent of UPDB. For example, the Cache County Memory and Health Study was launched in 1995 to study factors related to dementia and Alzheimer's disease risk. Participants were 65 and older at enrollment and were from a single county in Northern Utah. With the appropriate approvals, all were linked to UPDB. This linkage provided an opportunity to open up new life course studies of dementia and Alzheimer's disease (Norton et al., 2010, 2011, 2016).

In the end, the diversity of data sources and the annual updates of many of the data sources has created a resource in UPDB that includes nearly all of the residents of Utah. An assessment of the number of people alive and living in Utah in 2010 based on US census estimates shows close agreement with those represented in the UPDB.

3 CONCEPTUAL MODEL

A fundamental goal of the UPDB is to preserve the integrity of the data in the form in which it was received and yet create a set of unique individuals which can easily be used for record linking, statistical analysis and pedigree construction. Therefore, each dataset added to UPDB and all individuals listed on the records from those distinct datasets are assigned a unique dataset-specific identification number while relationships are created between individuals, such as a husband and wife on a marriage record or parent and child on a birth certificate. Information that is unique to each data source and time period is stored together in a separate dataset such as the manner of death on a death certificate or birth weight on a birth certificate. Major format changes to vital records with different information collected result in separate datasets. Personal information that is common across many data sources and is used in the matching process, is stored together in other datasets, including, but not limited to, demographic information, names, places, addresses, and relationships.

Initially, as each individual is loaded into UPDB, they exist with all their original information (archival information) and they also exist as a "composite" person, but the composite person only reflects the data originally received. After at least two or more persons are determined to be the same individual via the linking process, to facilitate further record linking and analysis of the data, the unique "composite" person record is re-created for that individual. The composite person is created by using a rules-based program which evaluates discrepant information. So, this rules-based program is only used after a link between two persons has been established to determine the most accurate and current name, demographic and relationship information of an individual based on the frequency and source of information. The objective is to create a person-oriented data structure where the person-specific information is selected in order to construct as complete as possible the life history of the individual from the many streams of data representing that individual. There are instances where the source data comprising the individual's life seems correct based on information at a given time but are deemed in error (or at least some portion) based on new information that subsequently comes to light as new records are added and linked. In this way, the person-oriented model is dynamic as new data are added to UPDB.

The durability, sustainability, and success of UPDB can be attributed in large measure to several factors outside the UPDB structure. First, complex and large linked databases such as UPDB are understandably expensive to build and maintain. In this instance, the Huntsman Cancer Institute has provided support to the Pedigree and Population Shared Resource (PPR) since the mid-1990s. This institutional foundation has given the PPR staff the stability it needs to engage in rational planning and support growth. This institutional basis has also been supplemented by the University of Utah beyond that provided by the Huntsman Cancer Institute. This funding model has been essential for the growth and the quality of UPDB. It also permits individual investigators to propose studies using UPDB data where the infrastructure costs have been largely paid by the institution (through philanthropic giving from the Huntsman Cancer Foundation and returned overhead to the University of Utah from extramural grants). Accordingly, research grants can accommodate the project-specific costs associated with PPR expenses through support from federal agencies. This has been a successful model given the large number of extramurally funded grants awarded to investigators using UPDB data.

A more subtle but important aspect of UPDB's success relates to the relatively small size of Utah's population and institutions. Utah's small population at the outset in the mid-1970s likely contributed to its inauguration. The volume of data was more manageable and the ability of the principal institutions to interact was conducive to creating a collaborative atmosphere between the key institutions (the Genealogical Society of Utah, the University of Utah, and the Utah Department of Health). The geographic proximity of these institutions contributed to negotiations and agreements that would likely have been more problematic in much larger states.

The growth and evolution of investigators and topics reliant on UPDB can in part be attributed to the catalyzing effects of big data on team science (Sellers et al., 2006; Shah, Pico, & Freedman, 2016; Stokols, Misra, Moser, Hall, & Taylor, 2008). The diversity and quality of UPDB data that is curated and made available has served to induce large and ambitious projects that require investigators from multiple disciplines. This has created teams that often combine medical, population and social sciences. Such multidisciplinary efforts generally serve to make the science stronger and have served to make UPDB essential to the larger research mission of the University of Utah.

4 RECORD LINKING

4.1 OVERVIEW

The linking process is fundamental to the core purposes of the UPDB, its utility, the representation of the diverse data sets it comprises, and the structure and scope of the pedigrees it contains. The objective is to identify efficiently the same individuals across millions of records historically as well as with each scheduled update of new records. The "composite" person is created using available identifying information including (when available) full name, birth date, death date, addresses, phone numbers, place of birth or death, encrypted Social Security Number, and names and specific relationships of family members.

Linking is accomplished primarily using probabilistic techniques supplemented by deterministic linking and manual linking as a result of manual review. The probabilistic linking software used for UPDB has evolved over time, from a command line program called Automatch using probabilistic linkage techniques based on Howard Newcombe's seminal work (Fair, Lalonde, & Newcombe, 1991; Newcombe, 1969; Newcombe, Kennedy, Axford, & James, 1959) to the current linking software called QualityStage, IBM's Websphere Information Integration Solution™ family of tools and applications (IBM, Armonk, NY, USA). QualityStage draws on information theory and advanced pattern recognition features to provide the highest level of automation for standardization and matching (Duvall et al., 2012).

For some data sources, the information is insufficient to use with probabilistic linking techniques. For example, Ambulatory surgery records may not provide names but only contain encrypted Social Security number, birth date, gender and ZIP code; in this situation, deterministic linking is used. Also, if a child on a birth certificate is linked to his or her own death certificate using probabilistic methods, then the parents listed can be linked with assurance using deterministic methods with only their names. The principles of using a combination of probabilistic and deterministic linking techniques with systematic validation supplemented with hand edits as needed has remained constant for UPDB record linkage throughout the database's existence. Validation processes external to QualityStage are used which may result in manual review of potential links.

The process of UPDB record linkage begins with the receipt of new records that are scheduled to be transferred to the UPDB team every six months or every year from the data contributors. The fact that UPDB adds contemporary data (perhaps with a one to one and a half year lag from the date of the record, a lag induced by internal data processing required by the data contributors to achieve their original mandate) is a central strength of the UPDB for studying contemporary outcomes which may be linked to more distant historical circumstances. This tempo of creating up-to-date information on living individuals within UPDB and building their life histories is attractive to researchers from a range of disciplines who are often focused on past causal events that may affect current responses (e.g., diagnosed with COVID-19).

4.2 INFORMATION STANDARDIZATION

All source records are securely transferred and loaded onto UPDB servers. Individuals in these data sets are represented by records created in UPDB, assigned an ID specific to the source (which is distinct from the person-level ID for the composite person), and the variables in these records are standardized according to UPDB protocol. For example, Social Security numbers are encrypted, punctuation and spaces are removed from within names, street addresses are standardized to match the US Postal Service standards (e.g., Street is abbreviated to ST) while cities, counties, states and countries are matched against a UPDB-specific dictionary to remove common spelling errors and abbreviations. When a record is received with multiple individuals identified with their family relationships, such as on a birth certificate (child, mother, father), a distinct record for each individual on this record is created in UPDB, assigned an ID number and these genetic or marital relationships between the individuals are indicated. Each individual record exists with all the information that was received initially in the source (archive) record; some of this information is also maintained in the record of the "composite" individual or Person Record.

The information that is used for record linking is selected from the tables of standardized information for the "composite" person and a single file is created. This file contains sex, names (original and Soundex), birth dates, death dates, birth place, death place, whether the individual is a twin/triplet, encrypted Social Security numbers, current address, and phone number of the individual. Also included are parent's name and their encrypted Social Security number as well as spouse's name, birth date, encrypted Social Security number, marriage date and birth date. Multiple records will exist for an individual with multiple spouses. For linking census records, information on the four eldest children are also included. All of these fields are available for use in the QualityStage program. Additional information can be used for validation, such as previous addresses for one person matching the current address of a potential link.

As additional records are loaded and linked, the validity and quantity of the information on a given person increases. In addition, social (e.g., marriage) and genetic (offspring) relationships are added and verified as multiple records containing relationship information are added and linked. An example of this arises in the case of names appearing on birth and death certificates. Parents who are listed on a death certificate can also be listed as parents on the decedent's birth certificates. When children on birth certificates are linked to their own death certificate, the two sightings of the parents (once on each vital record) are evaluated and linked if possible. In this example, a relationship may change from being identified as a birth parent relationship on a death certificate to an adoptive parent.

The resulting data are stored in tables in a relational database that cover a range of concepts or domains. These domains are numerous and include relationships (e.g., ego-mother-father), demographic features, medical diagnoses, insurance claims, birth/death details, residential history, and follow-up information. This domain-oriented data structure may draw on information from a variety of sources or they may be based on a single source. Tailored datasets approved for analysis will be created from multiple domains.

4.3 GENERAL LINKING STRATEGY

Potential links are initially created based on exact matching on one or more fields. These fields are called blocking variables or fields which create a subset of records that satisfy this exact matching. For example, the combination of encrypted Social Security number and birth year or the combination of first name, last name and birth year that exactly match across two records would be assessed and appear in the subset. Many different blocking combinations are used to account for instances where only one character may differ in a name or a birth year may differ by a single year.

Within sets of blocked records, statistical weight (affecting soundex) are then calculated for fields with sufficient variability (e.g., last name, first name, birth place, mother's maiden name) and used to measure the contribution of these fields to the probability of matching two records accurately. These weights are an extension of the Fellegi and Sunter algorithm (Fellegi & Sunter, 1969) developed by Jaro (Jaro, 1995; see also DuVall, Kerber, & Thomas, 2010) and derived from probabilities that utilize the frequency of the distinct values in the field which are generated by QualityStage. These field-specific agreement weights are based on the probability that the field agrees given that the records are true matches (m -probability) and the probability that the field agrees given that the records are

not true matches (u-probability). The combination of the m- and u-probabilities form the basis of the weight assigned to the matching for a given field. If the fields do not match, a disagreement weight is assigned dependent on parameters that define the likelihood of a mismatch given that the two records are true matches; therefore, knowledge of the data and their quality can be incorporated into the weights. A positive weight for comparison of mismatched variables may be assigned by evaluating the similarity of the two strings using an algorithm that is based on information theory principles. A composite weight is computed by summing the distinct (dis)agreement component weights of each variable comparison. Threshold values are used to classify a "good" link if the composite weight is above a threshold value, a nonmatch if it is below lower threshold, and undecided otherwise (these undergo manual review or further validation with other variables). A series of passes over the data are performed using different combinations of blocking and matching fields. Relationships between individuals can also be utilized for blocking. Several illustrative examples include:

- The first name on one record may be blocked with the middle name from another record; the last name on one record of a woman may be blocked with the spouse's last name from another record where the woman is recorded with only her maiden name.
- To address potential keying errors for date fields, birth day on record A is blocked with birth month on record B and birth month on record A is blocked with birth day on record B.
- Relationships that exist between parent and children as well as spouses allow for blocking on their names.
- The US Census of Utah records have been linked to genealogy records using blocking on parent's name and four of their eldest children's names or with just a single relationship between child – mother, child – father or husband – wife.

Each set of candidate links are processed through a set of validation checks before being incorporated into the UPDB. These validation programs identify links that may need additional attention and manual review. The set of links to be reviewed indicate the check that generated the need for additional attention and the source of the inconsistency. These checks include general logical inconsistencies, such as children born before parents, born to implausibly young parents, having two birth/death certificates, or different birth places. Often information from family members may be used to help validate a questionable link. For instance, if encrypted Social Security numbers do not match for an individual, they are compared with those of relatives (parents, spouse, children) for the case where the encrypted Social Security numbers is used on a record but is not that of the individual in question. The same process may be performed that involve mismatched addresses and phone numbers.

Finally, records identified as valid matches, compiled from QualityStage based on composite scores, validation programs, and human review, are then processed through the "composite" person creation program and incorporated into UPDB. On a monthly basis, additional validation programs are run and assigned a priority value with the highest priority given to the most egregious logical inconsistencies and resolved. To provide an example, due to timing of the processing of links, imagine Individual A who has minimal information (example, name and birth date only) and who may link to two different individuals (B and C). The validation program which assesses A and B as well as A and C separately may not identify any problems. However, after the composite person is created from information from Individual A, B and C, the logical inconsistencies may arise such that the composite person now has multiple parents with different birth dates indicating an incorrect link. There are additional procedures in place to assess this problem of multiple records linked during a single linking run, however with multiple linkers and the lag time of creating the "composite" person, some of these links can only be caught during the monthly validation checks.

When invalid links are discovered, these links are broken and then new composite person records are created with a re-assessment of the best information to retain. There is also a process to permanently reject links so that they do not happen again with a new linking process by adding the IDs to a table that is checked every time a new link is processed. If new information comes to light, the rejected link can be removed from the table. This is a process that always involves human intervention.

5 PRIVACY AND CONFIDENTIALITY

Our principal concern regarding use of linked datasets in UPDB is protecting identities of individuals in these data. To establish some basic definitions, privacy refers to an individual's ability to control information about him/herself while confidentiality is the obligation of a second party to not reveal private information about an individual to a third party without the permission of the person concerned (Wylie & Mineau, 2003).

When individuals agree to participate in research studies or when UPDB data are provided to researchers, it is with the understanding that the information will only be used to advance research and will be kept confidential. Only the minimum data necessary to conduct the research is provided. Strategies such as removing explicit identifiers, e.g. name, full birth date, street address, and Social Security number, have been used to ensure confidentiality before releasing information to researchers.

Even when these measures have been implemented, potential re-identification methods could be used, such as matching to other databases or by looking at unique characteristics found in the fields of the database itself resulting in possible deductive disclosure of the identities of the individuals represented. Even when current methods of protecting identifying information from researchers are employed, there is still some risk to privacy and confidentiality when linking and sharing health information for research (Gymrek, McGuire, Golan, Halperin, & Erlich, 2013).

Using identifiers in UPDB is designed to optimize matching individuals across data sets, whether linking is being conducted with historical or contemporary data, so other approaches to protect confidentiality need to be employed. Because state regulations regarding individually identifying information may differ and because federal regulations and requirements will vary according to type of information and its use, the protections for privacy and confidentiality have to be tailored in different areas to comply with those regulations.

5.1 RESOURCE FOR GENETIC AND EPIDEMIOLOGIC RESEARCH (RGE)

Access to UPDB data is regulated by the Utah Resource for Genetic and Epidemiologic Research (RGE). The RGE was created by an Executive Order of the Governor of Utah on July 14, 1982. Relying on enabling statutes in state health code, the RGE was established as a "data resource for the collection, storage, study, and dissemination of medical and related information" to operate "for the purpose of reducing morbidity or mortality, or for the purpose of evaluating and improving the quality of hospital and medical care". Originally administered under the direction and supervision of the Utah Department of Health, the RGE was transferred to the University of Utah by a second Executive Order in 1986. RGE is the legal custodian for the data contained within the UPDB and is responsible for developing and maintaining contractual agreements with organizations that contribute data to the UPDB or that links records to the UPDB.

Each project requesting access to data from the UPDB or linked electronic medical records applies to RGE for review. Applications are reviewed by the RGE Committee, which includes representatives from the university faculty with expertise in several disciplines including demography, genetics, public health and epidemiology, as well as representatives from each of the data contributors. Each data contributor has the right to veto the use of its own data if the representative determines the proposal describes an inappropriate use of its data. In practice, representatives of the data contributors rarely exercise their veto power because most applications can be revised to address concerns. All projects are also required to obtain approval by the appropriate Institutional Review Board(s) and Privacy Board(s) before access to data is granted.

RGE has the responsibility to protect the sensitive confidential information in UPDB. The RGE requires that users with access to data sign the RGE Confidentiality and Data Use Agreement. Each user on a project is also required to disclose any relationship with a for-profit company that might have an interest in the research being conducted with UPDB data. Relationships with for-profit companies are not prohibited, but require an assurance that data will be protected. The RGE Committee evaluates the data security for each location in which UPDB data will be stored. Finally, before any research is published, investigators must submit manuscripts to RGE for review, which includes scrutiny of any potentially identifying data presented for publication.

5.2 DATA SHARING AND RELATIONSHIPS WITH DATA CONTRIBUTORS

Important issues arise when collaborating with agencies who contribute data to the UPDB. For UPDB to exist, it requires the full participation of numerous organizations interested in advancing research and willing to share data for the purposes of research. The University of Utah is the steward of the data comprising UPDB but these data are not owned by the University of Utah. Formal agreements with the contributing agencies to facilitate long-term sustainability of using linked datasets for research in UPDB have been addressed and involve addressing the following issues:

1. Most datasets in UPDB were not collected specifically for research but are allowed for research use with consideration of privacy and confidentiality concerns by the data contributor.
2. Data collected by investigators for administrative purposes, research (including biospecimen data), and from high-risk clinics are linked to UPDB, but any diagnostic, relationship or residential information cannot be released until permission is given by the investigator who provided the data to UPDB to be used by other investigators with appropriate IRB approval.
3. RGE works with data contributors to carefully describe to data users the data contributors' authority for allowing research use of the contributors' data linked to other datasets.
4. RGE and PPR staff negotiate agreements with data contributors regarding data security measures and the methods used to protect the confidentiality of these data.
5. When scientific discoveries are made, intellectual property needs to be clarified by formal agreements between institutions since these scientific advancements are based on links between data sets which represent new and synergistic information.

We illustrate these principles with an example that relates to an NIH grant directed by Dr. Ken Smith. In that study, Medicare claims data (Principle #1) were requested to allow age-eligible individuals in UPDB to be matched to their Medicare records (Principle #2). The University of Utah owns the links but not the Medicare data themselves which RGE explains to users (Principle #3). In this sense, researchers may function in ways that are similar to data contributors since they may (1) contribute the links outright to the research resource or (2) maintain control over future use, as institutional data contributors do, while establishing the necessary data security and privacy protections. For the latter, Medicare records must remain on a single secure server in a manner compliant with the requirements of the Centers for Medicaid and Medicare Services, the federal agency that collects and allows approved release of Medicare files (Principle #4). Several studies have used these Medicare data linked to UPDB (Hanson, Horn, Rasmussen, Hoffman, & Smith, 2017; Hanson, Smith, & Zimmer, 2015; Hollingshaus et al., 2016; Wirostko et al., 2016), publications that acknowledge the value and approved access to the Medicare data (Principle #5).

5.3 UPDB AS A RESEARCH RESOURCE

Certain kinds of research infrastructures, secure data centers, or statistical coordinating centers often link data sets for research use such as is done with UPDB. These research entities generally hold identifying data, link records and then provide approved de-identified data or limited data sets to investigators. Such entities have policies associated with the release of these linked data information. Like other such research entities, UPDB relies on the following RGE policies and procedures:

1. To create "minimum-necessary" (often de-identified or limited) data sets to be released for research use.
2. To preclude researchers from linking to other data resources (without approval) to prevent disclosure of individual information outside the scope of the original research agreement (Kohane & Altman, 2005; Winickoff, 2006).
3. To develop confidentiality agreements with users/investigators that require users not attempt to re-identify individuals and will disclose any breeches of confidentiality to RGE.
4. To develop methods for contacting and recruiting individuals for participation in a research protocol (Wylie & Mineau, 2003).

There are several models for creating research resources. The concept of the "Charitable Trust" has been suggested from the field of genomics biobanks (Winickoff & Winickoff, 2003). This means that academic medical centers might (and often) elect to transfer blood, tissue, and medical data to private biobanks in an exchange for access for research and equity. With this Charitable Trust approach,

when a person agrees to donate tissue, the trust is the steward of the tissue and is obligated to ensure protection of the donated tissue. Maintaining the viability and sustainability of the Trust is a challenge with this strategy, for participants, universities and for-profit organizations (Master, Campo-Engelstein, & Caulfield, 2015; Turner, Dallaire-Fortier, & Murtagh, 2013).

Others have proposed disease registries (e.g., statewide cancer registries) that could use a national system that separates information under the control of three groups, including the Disease Registry, a Population Registry (a trusted agency that maintains the personal identifying information) and an Identifier Translation Agency (another trusted third party that has the key to translate the unique identifier assigned by the Population Registry and by the disease registry) (Churches, 2003). This strategy would ensure safety but makes linking data more cumbersome, reduces the effectiveness of identity matching, increases costs, and imposes added costs for conducting research.

Some agencies, institutes and centers may provide infrastructure for record linking activities. Major data providers allow access to confidential data at secure data centers for approved users who have achieved security clearances. Gaining access to confidential micro-data, such as those held by the US Census and National Center for Health Statistics through the Federal Statistical Research Data Centers, represents an important example of this strategy. There are also university data centers that operate secure computing systems that have policies that allow researchers to acquire, maintain, and analyze restricted-use data.

In Utah, RGE is a resource that has addressed these issues for decades. It is a dynamic institution whose policies and procedures address increasing complexities related to managing and linking individual-identifying records for research use, and embodies elements of all three approaches: RGE controls access to the data as they are provided to UPDB through agreements with data contributors. UPDB encrypts Social Security Numbers when working with the statewide healthcare facilities and claims records as required by the Utah Department of Health. Medical record numbers are replaced with random unique IDs by University of Utah Health and Intermountain Healthcare before linking their electronic health records to the UPDB.

6 FEATURES OF UPDB THAT FACILITATE THE INTEGRATION OF GENETICS AND DEMOGRAPHY

There are substantial opportunities afforded to researchers using UPDB. Every year, the number of generations represented increases such that the detection of familial aggregation of diseases and outcomes improves. UPDB provides the capacity to identify multigenerational pedigrees with significant excess prevalence of specific conditions and to then enroll them for genetic, outcomes or other population-based studies. Indeed, deeper and larger comprehensive genealogies enhance the likelihood of gene discoveries. Because records from many data sets are linked together, UPDB is able to combine, confirm, or improve information at the individual level. Creating and maintaining a database similar to the UPDB would require resources beyond the scope of any single research project. While the genealogy records in UPDB may appear to be similar to those available through web-based genealogical databases, they are not since those sources generally only represent primary source data for use by individuals doing their own genealogies. With the use of vital records and driver license records, UPDB is a statewide resource and individuals born and living in Utah are more comprehensively represented. UPDB supports the larger goal of treating the entire state as a platform for genetic, population and outcomes research.

The value of the UPDB comes, in large measure, to the synergies arising from the number, time coverage and diversity of records added to the resource. A number of new sources of records and infrastructure developments exist that will expand the data collection needed to improve the utility of the UPDB.

1. **Data Coverage Expansion.** In addition to the annual updates of data sources described previously, UPDB continues to add other records such as historic birth certificate data and historic Census records (1950) as they become available.

2. Environmental and Geo-Spatial Capabilities. Expansion of the linkage of georeferenced environmental-geographic-socioeconomic data to UPDB facilitates environmental epidemiology, health services research, gene-environment interaction research and studies of social disparities.
3. Utah Genome Project and the Center for Genomic Medicine at the University of Utah. The Utah Genome Project (UGP) is a large-scale, multi-year initiative to advance better disease prevention, diagnosis, and treatment methods through discovery of genetic signatures for human diseases and response to drug therapies. UGP supports projects that collect biologic samples and sequence DNA identified through UPDB families with excess burden of disease; diseases targeted for support from UGP have significant public health impact and are deemed to be scientifically feasible targets for analysis. The UGP is a component of the Center for Genomic Medicine (CGM) which supports additional analytic and translation objectives along with data access to the UPDB.
4. Visualization and a New Pedigree construction. Important enhancements to UPDB's Kinship Analysis Tools (KAT) (Kerber, 1995; Kerber, O'Brien, Smith, & Cawthon, 2001) are being made. The use of visualization and network-based tools takes a set of individuals of interest, provided by a researcher, and traverses the full extent of genealogies in the UPDB, connecting these individuals through all existing family relationships, whether close or distant relatives, and creating a comprehensive multi-lineage pedigree. These programs are fully scalable and will fill a current gap in describing family structure, social networks and provide depictions of complex pedigrees necessary for sophisticated genetic analyses. Additional tools are being developed for the analysis of genetic heterogeneity and gene-environment interactions. (Hanson et al., 2020)
5. UPDB Limited (UPDB-L) Query Tool. Potential investigators may access large subsets of data from the UPDB through the online UPDB Limited Query Tool. Version 1.0 was released in 2009 and provided access to all death and birth certificates, Inpatient Hospital Claims and Ambulatory Surgery Claims, statewide cancer diagnoses, geographic and demographic information, along with data on familial relationships and pedigrees. Plans are in place to expand the tool to include emergency department claims, All Payer Claims Database and University of Utah health data.
6. UPDB Linkages to Biospecimens and Clinical Measures. The UPDB has benefitted from linkages to biospecimens. In working with the Huntsman Cancer Institute's Research Informatics Shared Resource and the Biospecimen and Molecular Pathology, UPDB is linked to clinically annotated biobanking, histology services, and molecular diagnostics. Since UPDB is also linked to the records of Intermountain Healthcare and the University of Utah Hospitals and Clinics, these sources also hold clinical data that arise as a matter of patient care; these clinical data are only transferred to researchers from the respective clinical enterprise data warehouses when projects are approved.

7 CONCLUSION

The UPDB offers demographers, historians, geneticists, oncologists, physicians, epidemiologists, and other social scientists an unparalleled data resource from which to launch the next generation of studies that rely on data heretofore unavailable, including the prospect of novel data types such as full genome and exome sequencing. Access to these novel data joined with the depth of information from the UPDB make it an extremely attractive research resource for both investigators and their trainees and students and have contributed to the success and popularity of the UPDB. Moreover, given the sensitivity of the data (spanning family relationships, linkages to DNA and biobanks, geospatial markers), users of the UPDB receive the protections and oversight of the Utah Resource for Genetic and Epidemiologic Research (RGE) that have been in place for decades and permit responsible and ethical use of the data while protecting the identities of the individuals whose data are the basis for the research undertaken. Over time, the UPDB has grown in terms of the number of individuals and families represented as well as the diversity of data sources. This growth, with the proper privacy protections, portend continued use of UPDB across a range of topics and disciplines. Nonetheless, the stakes remain high when managing such large volumes of data. The structure of the UPDB requires that it continues to earn the trust and confidence of the public, state government representatives, and

the data contributors. In the end, UPDB represents a valuable data resource which scientists can use to test next-generation hypotheses.

REFERENCES

- Adams, J., Lam, D. A., Hermalin, A. I., & Smouse P. E. (Eds.) (1990). *Convergent issues in genetics and demography*. New York: Oxford University Press.
- Bean, L. L., May, D. L., & Skolnick, M. (1978). The Mormon historical demography project. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 11(1), 45–53. doi: [10.1080/01615440.1978.9955216](https://doi.org/10.1080/01615440.1978.9955216)
- Casey, J. A., Schwartz, B. S., Stewart, W. F., & Adler, N. E. (2016). Using electronic health records for population health research: A review of methods and applications. *Annual Review of Public Health*, 37, 61–81. doi: [10.1146/annurev-publhealth-032315-021353](https://doi.org/10.1146/annurev-publhealth-032315-021353)
- Cawthon, R. M., Smith, K. R., O'Brien, E., Sivatchenko, A., & Kerber, R. A. (2003). Association between telomere length in blood and mortality in people aged 60 years or older. *The Lancet*, 361(9355), 393–395. doi: [10.1016/S0140-6736\(03\)12384-7](https://doi.org/10.1016/S0140-6736(03)12384-7)
- Chernenko, A., Meeks, H., & Smith, K. R. (2019). Examining validity of body mass index calculated using height and weight data from the US driver license. *BMC Public Health*, 19, 100. doi: [10.1186/s12889-019-6391-3](https://doi.org/10.1186/s12889-019-6391-3)
- Churches, T. (2003). A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers. *BMC Medical Research Methodology*, 3, 1. doi: [10.1186/1471-2288-3-1](https://doi.org/10.1186/1471-2288-3-1)
- DuVall, S. L., Fraser, A. M., Rowe, K., Thomas, A., & Mineau, G. P. (2012). Evaluation of record linkage between a large healthcare provider and the Utah Population Database. *Journal of the American Medical Informatics Association*, 19(e1), e54–59. doi: [10.1136/amiajnl-2011-000335](https://doi.org/10.1136/amiajnl-2011-000335)
- DuVall, S. L., Kerber, R. A., & Thomas, A. (2010). Extending the Fellegi–Sunter probabilistic record linkage method for approximate field comparators. *Journal of Biomedical Informatics*, 43(1), 24–30. doi: [10.1016/j.jbi.2009.08.004](https://doi.org/10.1016/j.jbi.2009.08.004)
- Fair, M. E., Lalonde, P., & Newcombe, H. B. (1991). Application of exact ODDS for partial agreements of names in record linkage. *Computers and Biomedical Research*, 24(1), 58–71. doi: [10.1016/0010-4809\(91\)90013-m](https://doi.org/10.1016/0010-4809(91)90013-m)
- Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183–1210. doi: [10.1080/01621459.1969.10501049](https://doi.org/10.1080/01621459.1969.10501049)
- Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science*, 339(6117), 321–324. doi: [10.1126/science.1229566](https://doi.org/10.1126/science.1229566)
- Hammad, I. A., Meeks, H., Fraser, A., Theilen, L. H., Esplin, M. S., Smith, K. R., & Varner, M. W. (2020). Risks of cause-specific mortality in offspring of pregnancies complicated by hypertensive disease of pregnancy. *American Journal of Obstetrics and Gynecology*, 222(1), 75.e1–75.e9. doi: [10.1016/j.ajog.2019.07.024](https://doi.org/10.1016/j.ajog.2019.07.024)
- Hanson, H. A., Horn, K. P., Rasmussen, K. M., Hoffman, J. M., & Smith, K. R. (2017). Is cancer protective for subsequent Alzheimer's disease risk? Evidence from the Utah Population Database. *The Journals of Gerontology: Series B*, 72(6), 1032–1043. doi: [10.1093/geronb/gbw040](https://doi.org/10.1093/geronb/gbw040)
- Hanson, H. A., Leiser, C. L., Madsen, M. J., Gardner, J., Knight, S., Cessna, M., . . . Camp, N. J. (2020). Family study designs informed by tumor heterogeneity and multi-cancer pleiotropies: The power of the Utah Population Database. *Cancer Epidemiology, Biomarkers & Prevention*, 29(4), 807–815. doi: [10.1158/1055-9965.EPI-19-0912](https://doi.org/10.1158/1055-9965.EPI-19-0912)
- Hanson, H. A., Smith, K. R., & Zimmer, Z. (2015). Reproductive history and later-life comorbidity trajectories: A medicare-linked cohort study from the Utah Population Database. *Demography*, 52(6), 2021–2049. doi: [10.1007/s13524-015-0439-5](https://doi.org/10.1007/s13524-015-0439-5)
- Hollingshaus, M. S. (2015). *Seeds of sorrow: A life-course approach to early-life parental death and later-life suicide and behavioral health risk* (Doctoral dissertation). Utah: University of Utah. Retrieved from <https://collections.lib.utah.edu/ark:/87278/s6060q8v>
- Hollingshaus, M. S., Coon, H., Crowell, S. E., Gray, D. D., Hanson, H. A., Pimentel, R., & Smith, K. R. (2016). Differential vulnerability to early-life parental death: The moderating effects of family suicide history on risks for major depression and substance abuse in later life. *Biodemography and Social Biology*, 62(1), 105–125. doi: [10.1080/19485565.2016.1138395](https://doi.org/10.1080/19485565.2016.1138395)

- Hurdle, J. F., Smith, K. R., & Mineau, G. P. (2013). Mining electronic health records: An additional perspective. *Nature Reviews Genetics*, *14*, 75. doi: [10.1038/nrg3208-c1](https://doi.org/10.1038/nrg3208-c1)
- Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine*, *14*(5–7), 491–498. doi: [10.1002/sim.4780140510](https://doi.org/10.1002/sim.4780140510)
- Kerber, R. A. (1995). Method for calculating risk associated with family history of a disease. *Genetic Epidemiology*, *12*(3), 291–301. doi: [10.1002/gepi.1370120306](https://doi.org/10.1002/gepi.1370120306)
- Kerber, R. A., O'Brien, E., Smith, K. R., & Cawthon, R. M. (2001). Familial excess longevity in Utah genealogies. *The Journals of Gerontology: Series A*, *56*(3), B130–139. doi: [10.1093/gerona/56.3.b130](https://doi.org/10.1093/gerona/56.3.b130)
- Kohane, I. S., & Altman, R. B. (2005). Health-information altruists — A potentially critical resource. *The New England Journal of Medicine*, *353*(19), 2074–2077. doi: [10.1056/NEJMs051220](https://doi.org/10.1056/NEJMs051220)
- Master, Z., Campo-Engelstein, L., & Caulfield, T. (2015). Scientists' perspectives on consent in the context of biobanking research. *European Journal of Human Genetics*, *23*(5), 569–574. doi: [10.1038/ejhg.2014.143](https://doi.org/10.1038/ejhg.2014.143)
- Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., . . . Skolnick, M. H. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. *Science*, *266*(5182), 66–71. doi: [10.1126/science.7545954](https://doi.org/10.1126/science.7545954)
- Neklason, D. W., Stevens, J., Boucher, K. M., Kerber, R. A., Matsunami, N., Barlow, J., . . . Burt, R. W. (2008). American founder mutation for attenuated familial adenomatous polyposis. *Clinical Gastroenterology and Hepatology*, *6*(1), 46–52. doi: [10.1016/j.cgh.2007.09.017](https://doi.org/10.1016/j.cgh.2007.09.017)
- Newcombe, H. B. (1969). The use of medical record linkage for population and genetic studies. *Methods of Information in Medicine*, *8*(1), 7–11.
- Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records. *Science*, *130*(3381), 954–959. doi: [10.1126/science.130.3381.954](https://doi.org/10.1126/science.130.3381.954)
- Norton, M. C., Fauth, E., Clark, C. J., Hatch, D., Greene, D., Pfister, R., . . . Smith, K. R. (2016). Family member deaths across adulthood predict Alzheimer's disease risk: The Cache County Study. *International Journal of Geriatric Psychiatry*, *31*(3), 256–263. doi: [10.1002/gps.4319](https://doi.org/10.1002/gps.4319)
- Norton, M. C., Smith, K. R., Østbye, T., Tschanz, J. T., Corcoran, C., Schwartz, S., . . . Welsh-Bohmer, K. A. (2010). Greater risk of dementia when spouse has dementia? The Cache County Study. *Journal of the American Geriatrics Society*, *58*(5), 895–900. doi: [10.1111/j.1532-5415.2010.02806.x](https://doi.org/10.1111/j.1532-5415.2010.02806.x)
- Norton, M. C., Smith, K. R., Østbye, T., Tschanz, J. T., Schwartz, S., Corcoran, C., . . . Welsh-Bohmer, K. A. (2011). Early parental death and remarriage of widowed parents as risk factors for Alzheimer disease: The Cache County study. *The American Journal of Geriatric Psychiatry*, *19*(9), 814–824. doi: [10.1097/JGP.0b013e3182011b38](https://doi.org/10.1097/JGP.0b013e3182011b38)
- Sellers, T. A., Caporaso, N., Lapidus, S., Petersen, G. M., & Trent, J. (2006). Opportunities and barriers in the age of team science: Strategies for success. *Cancer Causes & Control*, *17*, 229–237. doi: [10.1007/s10552-005-0546-5](https://doi.org/10.1007/s10552-005-0546-5)
- Shah, R., Pico, A. R., & Freedman, J. E. (2016). Translational epidemiology: Entering a brave new world of team science. *Circulation Research*, *119*(10), 1060–1062. doi: [10.1161/CIRCRESAHA.116.309881](https://doi.org/10.1161/CIRCRESAHA.116.309881)
- Skolnick, M., Bean, L. L., Dintelman, S. M., & Mineau, G. (1979). A computerized family history data base system. *Sociology and Social Research*, *63*(3), 506–523.
- Skolnick, M., Bean, L., May, D., Arbon, V., De Nevers, K., & Cartwright, P. (1978). Mormon demographic history I. Nuptiality and fertility of once-married couples. *Population Studies*, *32*(1), 5–19.
- Skolnick, M., Bishop, D., Carmelli, D., Gardner, E., Hadley, R., Hasstedt, S., . . . Smart, C. (1981). A population-based assessment of familial cancer risk in Utah Mormon genealogies. In F. E. Arrighi, P. N. Rao, & E. Stubblefield (Eds.), *Genes, chromosomes, and neoplasia* (477–500). New York: Raven Press
- Smith, K. R., Brown, B. B., Yamada, I., Kowaleski-Jones, L., Zick, C. D., & Fan, J. X. (2008). Walkability and body mass index: Density, design, and new diversity measures. *American Journal of Preventive Medicine*, *35*(3), 237–244. doi: [10.1016/j.amepre.2008.05.028](https://doi.org/10.1016/j.amepre.2008.05.028)
- Smith, K. R., Hanson, H. A., Mineau, G. P., & Buys, S. S. (2012). Effects of *BRCA1* and *BRCA2* mutations on female fertility. *Proceedings of the Royal Society B*, *279*(1732), 1389–1395. doi: [10.1098/rspb.2011.1697](https://doi.org/10.1098/rspb.2011.1697)
- Smith, K. R., Zick, C. D., Kowaleski-Jones, L., Brown, B. B., Fan, J. X., & Yamada, I. (2011). Effects of neighborhood walkability on healthy weight: Assessing selection and causal influences. *Social Science Research*, *40*(5), 1445–1455. doi: [10.1016/j.ssresearch.2011.04.009](https://doi.org/10.1016/j.ssresearch.2011.04.009)
- Song, X., & Campbell, C. D. (2017). Genealogical microdata and their significance for social science. *Annual Review of Sociology*, *43*(1), 75–99. doi: [10.1146/annurev-soc-073014-112157](https://doi.org/10.1146/annurev-soc-073014-112157)

- Stokols, D., Misra, S., Moser, R. P., Hall, K. L., & Taylor, B. K. (2008). The ecology of team science: Understanding contextual influences on transdisciplinary collaboration. *American Journal of Preventive Medicine*, 35(2 Suppl.), S96–S115. doi: [10.1016/j.amepre.2008.05.003](https://doi.org/10.1016/j.amepre.2008.05.003)
- Stroup, A. M., Herget, K. A., Hanson, H. A., Reed, D. L., Butler, J. T., Henry, K. A., . . . Smith, K. R. (2017). Baby Boomers and birth certificates: Early-life socioeconomic status and cancer risk in adulthood. *Cancer Epidemiology, Biomarkers & Prevention*, 26(1), 75–84. doi: [10.1158/1055-9965.EPI-16-0371](https://doi.org/10.1158/1055-9965.EPI-16-0371)
- Theilen, L. H., Fraser, A., Hollingshaus, M. S., Schliep, K. C., Varner, M. W., Smith, K. R., & Esplin, M. S. (2016). All-cause and cause-specific mortality after hypertensive disease of pregnancy. *Obstetrics & Gynecology*, 128(2), 238–244. doi: [10.1097/AOG.0000000000001534](https://doi.org/10.1097/AOG.0000000000001534)
- Theilen, L. H., Meeks, H., Fraser, A., Esplin, M. S., Smith, K. R., & Varner, M. W. (2018). Long-term mortality risk and life expectancy following recurrent hypertensive disease of pregnancy. *American Journal of Obstetrics and Gynecology*, 219(1), 107.e1–107.e6. doi: [10.1016/j.ajog.2018.04.002](https://doi.org/10.1016/j.ajog.2018.04.002)
- Turner, A., Dallaire-Fortier, C., & Murtagh, M. J. (2013). Biobank economics and the "Commercialization Problem". *Spontaneous Generations: A Journal for the History and Philosophy of Science*, 7(1), 69–80. doi: [10.4245/sponge.v7i1.19555](https://doi.org/10.4245/sponge.v7i1.19555)
- Williams, R. R., Skolnick, M., Carmelli, D., Maness, A. T., Hunt, S. C., Hasstedt, S., . . . Jones, R. K. (1979). Utah pedigree studies: Design and preliminary data for premature male CHD deaths. *Progress in Clinical and Biological Research*, 32, 711–729. Retrieved from PMID: [523491](https://pubmed.ncbi.nlm.nih.gov/523491/)
- Winickoff, D. E. (2006). Health-information altruists. *The New England Journal of Medicine*, 354(5), 530–531. doi: [10.1056/NEJMc053390](https://doi.org/10.1056/NEJMc053390)
- Winickoff, D. E., & Winickoff, R. N. (2003). The charitable trust as a model for genomic biobanks. *The New England Journal of Medicine*, 349(12), 1180–1184. doi: [10.1056/NEJMs030036](https://doi.org/10.1056/NEJMs030036)
- Wirotko, B. M., Curtin, K., Ritch, R., Thomas, S., Allen-Brady, K., Smith, K. R., . . . Allingham, R. R. (2016). Risk for exfoliation syndrome in women with pelvic organ prolapse : A Utah project on Exfoliation Syndrome (UPEXS) study. *JAMA Ophthalmology*, 134(11), 1255–1262. doi: [10.1001/jamaophthalmol.2016.3411](https://doi.org/10.1001/jamaophthalmol.2016.3411)
- Wylie, J. E., & Mineau, G. P. (2003). Biomedical databases: Protecting privacy and promoting research. *Trends in Biotechnology*, 21(3), 113–116 doi: [10.1016/S0167-7799\(02\)00039-2](https://doi.org/10.1016/S0167-7799(02)00039-2)
- Zick, C. D., Smith, K. R., Fan, J. X., Brown, B. B., Yamada, I., & Kowaleski-Jones, L. (2009). Running to the store? The relationship between neighborhood environments and the risk of obesity. *Social Science & Medicine*, 69(10), 1493–1500. doi: [10.1016/j.socscimed.2009.08.032](https://doi.org/10.1016/j.socscimed.2009.08.032)

HISTORICAL LIFE COURSE STUDIES
VOLUME 13 (2023), published 11-05-2023

The Development of Microhistorical Databases in Norway

A Historiography

Gunnar Thorvaldsen
UiT The Arctic University of Norway

Lars Holden
The Norwegian Computing Center

ABSTRACT

Norwegian work on microdata started out with the full count 1801 census and census and vital records from around the capital. Today, most census and ministerial records from 1801 until the mid-20th century have been scanned, transcriptions are being completed, much is encoded and made available via the websites of the Digital National Archives and UiT The Arctic University of Norway. This article complements a previous publication on empirical results from historical microdata. It is primarily organized by technical issues: digitization of source materials, encoding and standardization, building of the Historical Population Register for the period since 1800, record linkage and source criticism as well as GIS. Presently, partner institutions are building the Historical Population Register with prolonged support from the Norwegian Research Council. This will contain longitudinal records of the nine million persons who lived in Norway since 1800. The register increasingly makes it possible to follow the entire population. Unique personal IDs with corresponding URLs to the person page providing links to many sources introduce a new level of historical documentation. Cross-sectional and vital records are being interlinked with automatic and manual record linkage software. Longitudinal data is available for searching as timelines and in Intermediate Data Structure format from UiT The Arctic University and for searching at Histreg.no, which also caters for manual editing. We are well on the way to creating a database that can fill the void in the two centuries before the Central Population Register starts in 1964.

Keywords: Norway, Microdata, Censuses, Church records, Data processing, Population register, Record linkage

DOI article: <https://doi.org/10.51964/hlcs14315>

© 2023, Thorvaldsen, Holden

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

This article summarizes the history of the computerization of historical microdata in Norway. It is a technically oriented follow up to the historiography article highlighting the research accomplishments using such datasets originating in nominative historical sources (Sommerseth & Thorvaldsen, 2022). These are primarily full count nominative censuses, vital registers in church books and emigration records. The aim is to cover the main developments of national significance, including illustrative regional ones.

The article is primarily organized by technical issues: digitization of source materials, encoding and standardisation, building of the Historical Population Register for the period since 1800, record linkage as well as mapping and administrative borders. We start, however, with a brief overview of the organisation of the databases. Since many users, especially the less quantitatively oriented, are not so familiar with information technology we try to write without too many technicalities. On the other hand, some such details are necessary for the article to be useful for those who are building similar databases.

2 ORGANIZING THE HISTORICAL MICRO-DATA INFRASTRUCTURE

2.1 EARLY PROJECTS AT THE UNIVERSITIES OF BERGEN AND OSLO

The first historical Norwegian computer project was undertaken at the University of Bergen, aiming to computerize the full count 1801 census of Norway. Started in 1968, this was a joint venture between the Historical Institute, the National Archives and Statistics Norway. The major goal was easier access to one of the most central sources from "the old society." Then Norway was still largely rural with 800,000 inhabitants in the countryside, 80,000 town dwellers and 8,000 civil servants (Statistics Norway, 1980). Rather than converting the source material into numerical codes, it was wisely decided to enter the contents of the census verbatim. This source version soon became popular among archivists and genealogists, as well as social historians. The main guideline was to change as little as possible when transcribing the records onto rolls of punched paper tapes for the Univac mainframe. Student and soon researcher Jan Oldervoll not only supervised the punching but also developed software to print, sort and encode each census column.

A user-friendly Internet interface pioneered the 1801 census onto the World Wide Web and allowed open access to search for place names, personal names, and other census variables — unfortunately, this census contains no information on birthplace. Users were enabled to produce univariate and bivariate statistics, including the analysis of households classified according to the Hammel-Laslett system (Sommerseth, 2011; Statistics Norway, 1980). This 1801 census edition has been employed in a number of local history studies and was the central source for several master theses written in Bergen. They linked the census records interactively to baptisms, marriages and burials from the church books for 48 parishes spread across Norway. The most notable results pertain to the background and destiny of unwed mothers and social mortality differentials (Engelsen, 1983; Haavet, 1982). The Historical Institute at the University of Bergen also computerized parts of other censuses, emigration records, and ministerial records in cooperation with the Regional Archive in Bergen.

The microhistory project *Norwegian Social Development 1860 to 1900* was launched by the University of Oslo in 1971. Microhistory meant both history on the individual level and within a limited framework of time and space. Information about occupation, age and other variables must be available for each person on an individual level, not just aggregated for all residents of e.g., a municipality. In this way it would become possible to aggregate data to the desired level of analysis: families, the inhabitants of a census district, the entire municipality or larger regions. In turn, the life courses of groups of individuals were combined into collective biographies. The studies were thus both cross-sectional and longitudinal. The two selected locations were Ullensaker and the capital Kristiania (Langholm, 1974, 1976). These were the methodological and geographical contexts for studies of workers at the new Kværner Brug factory in Oslo and the radical popular movement started by Marcus Thrane in 1849. Mainframe computers were used both to organize the source material to find data about the individuals in order to describe featured groups statistically and to describe the whole population in the study areas.

The Ullensaker and Kristiania project developed software for transcribing the sources, proofreading, correcting errors, listing and sorting the material. This internally constructed program package HISO was also used to encode the datasets, while the standard statistical program package DDPP (Discrete Data

Program Package) of the mainframe DEC computer was used for aggregates. The computerized sources were the 1865, 1875 and 1900 censuses, parts of the emigrant records, as well as the church registers 1845 to 1875 for Ullensaker parish and the 1875 census for Kristiania which was the biggest component (78,000 individuals) anyway. The large costs of transferring the sources to the computer, could be justified with extensive research use by projects with many graduate students and collective supervision. More than 20 master theses were written based on these digitalizations. In addition, a number of students were inspired to study related themes for other locations. Central among the themes were social and geographical mobility, suitable themes based on the sources' information about occupation and birthplace, utilized with modified or full-fledged reconstitution methodology. Our knowledge increased especially about who moved to America or to the city, about connections between father's and son's occupations and about typical professional careers in the second half of the 19th century. The digitized source materials are stored in the National Archives and in the University of Tromsø.

2.2 THE NORWEGIAN HISTORICAL DATA CENTRE

The computerization of historical sources was transferred to the startup Norwegian Historical Data Centre (NHDC) at the UiT The Arctic University of Norway in Tromsø from 1978 onwards. Since 1985 it is a permanent body within the Faculty of Social Sciences and Humanities, serving researchers, teachers, students, and genealogists nationwide. The prime aim of the NHDC and its partners is a national population registry of the 18th and 19th centuries (cf. Section 4). The NHDC published printed books with verbatim transcriptions of the 1865, 1875, 1900 and 1910 census manuscripts as well as parish registers with alphabetical indexes. Also, digital versions following a national standard for data entry and data distribution were sent to the users (Nygaard, 1995). The encoded versions of the censuses were initially distributed on diskettes. In cooperation with the National Archives, the full count census transcriptions were expanded into nationwide datasets.

The transcription of parish registers was a more labor-intensive undertaking than the censuses, making present geographic coverage more limited, at 2/3, but higher in the 19th century. The handwriting in these sources is often Gothic, increasing the transcription difficulties significantly. The source material was made available as originals or xeroxed copies from the National or Regional Archives but now mainly as scanned versions via the Internet.

2.3 THE DIGITAL ARCHIVE

Through its Digital Archive, the National Archives present open access documents in digital formats. Many of the digitized sources contain nominative microdata in scanned or transcribed searchable formats, especially from censuses and church books, but also from emigration protocols, probate records and sailors' or military registers. The transcribed copies originate from the UiT The Arctic University of Norway, the National Archives' own transcription groups, international genealogical companies and volunteers. Scanned or transcribed sources can be chosen by period, topic, archive deposit or location of origin. When searching the transcribed versions, individuals can also be retrieved by names, gender, birthplace and -year/-date, status, type of event or role, event time and information about a relative (cf. the web interface page in Figure 1). Users may choose to search for exactly spelled names or for variants.¹ Certified researchers may upon application, program access to the records in the Digital Archive via an API gateway.

Registered users who log in to the Digital Archive are supposed to submit a correction notice if they believe, after inspecting the original, that a transcription is not correct. Users are warned that they may not suggest changes to what is written verbatim in the source. If users wish to enter additional information about a person, or to link information about the same person from several sources, they should use the online interface to the Historical Population Register (Histreg.no; see Section 4.2). The publication of transcribed church books through cooperation with the genealogical companies Ancestry, My Heritage and Family Search has resulted in a large increase in correction notices. Direct links between transcriptions and scanned images increasingly become available. Another development is that scanned printed and handwritten texts will progressively become searchable by employing OCR and handwriting recognition techniques.

¹ There is a version of the user interfaces in English at digitalarkivet.no/en, but some of the online help information is in Norwegian only.

Figure 1 Search page to retrieve individuals in the Digital Archive

The screenshot shows a search interface with two columns of input fields. The left column, titled 'Individual information', contains: 'First name:' with the value 'Ida Mathilde'; 'Last name:' with 'Hansd*'; 'Gender:' with a dropdown menu showing 'Female'; 'Role:' with an empty text box; 'Birth year:' with two dropdown menus, the first showing '1886' and the second showing '1886'; 'Birth date (mm-dd):' with a text box containing 'mm-dd'; and 'Place of birth:' with an empty text box. The right column, titled 'Event information', contains: 'Event year:' with 'From' and 'To' dropdown menus; 'Event date (mm-dd):' with a text box containing 'mm-dd'; and a 'Related person' section with fields for 'First name:', 'Last name:', 'Birth year:', and 'Role:', all of which are empty text boxes. At the top right, there is a 'Reset' button and a 'SEARCH' button with a magnifying glass icon.

When browsing scanned archive material in the Digital Archive, users will come across contents that is fully or partly restricted for use via the Internet. This applies to sensitive personal data about racial or ethnic origin, political opinion, religion, philosophical belief, trade union membership, criminal convictions or offenses. In addition, social security-like numbers can be blocked. However, place of birth, date of birth, citizenship, marital status, occupation, place of residence and place of employment are ordinarily not considered personal matters. In child welfare and adoption cases, the duty of confidentiality only expires after 100 years, and information in the state census manuscripts is obtained for use for statistical purposes only for 100 years according to the Statistics Act. The National Archives assume that the Personal Data Act does not provide protection for the deceased, and that sensitive personal data, can ordinarily be made available on the Internet when all the persons referred to have certainly died. Because it is difficult to establish that all persons referred to are dead, in many cases the 100-year rule applies anyway.

3 TRANSCRIPTION AND CODING

3.1 TRANSCRIPTION

The basic principle when transcribing censuses and other sources has been to copy the content as literally as possible. The chief method for achieving this aim has been to proofread the transcriptions, letting one assistant read from the digital transcription while another checks against the original source. We found this to be superior to double transcription because some transcribers tend to reproduce errors. And there are ergonomic reasons: proofreading is a less strenuous exercise than typing. As a last step we have an "acceptance control", where essential information (name, age) in 10% of the records are proofread to check that the proportion of errors is acceptable. Basically, this is a method to allow experienced transcribers to correct mistakes that were missed by less experienced colleagues during proofreading. Also, we sort the material to spot variable values with frequency one which indicate rare, erroneous cases and we program the computer to look for illogical combinations of variable values such as teenage widows. See the record linkage section 6.3 for techniques to spot conflicting information when we combine data from several sources.

However, there are exceptions to the verbatim transcription rule. We introduce a distinction between first names and surnames and between occupations and household positions. Census takers often used a special sign to indicate that the information is the same as from the previous person, which is replaced by the information itself. Unreadable handwriting is marked with double question marks, and illogical information with double exclamation marks. When in doubt about two different transcription alternatives, these can be separated by the masterspace "@". There are cases where the information is conflicting inside one census record, for instance when a "daughter" is marked as male in the gender field. Based on names and other information, the transcriber should correct what is obviously an error in the source and flag the correction in the comments field. Such checks are also built into the transcription apps which nowadays are constructed with standard database packages such as MS Access, rather than the "home brew" which was popular in past decades. To sum up: transcription rules are a compromise between creating a verbatim copy and enhancing the user-friendliness of the resulting database (Thorvaldsen et al., 2015).

3.2 HISTORICAL OCR PROJECTS

When starting to transcribe censuses and church records at the Norwegian Historical Data Centre in 1978 we aimed for a decentralized and low-tech solution: typing the content of the sources with ordinary typewriters. The typewriters were equipped with OCR-B balls, resulting in a special font that the rudimentary OCR readers at the time could recognize. This kind of OCR acted as a missing link after the heyday of the Hollerith punch card and before PCs became common. Anyone interested in this aspect of the history of computing should read *Travels in Computerland* (Schneider, 1974). We mailed our typed pages to a commercial company in Sweden who returned the contents as ASCII text files, for checking and correcting on a mainframe computer. The advent of PCs for transcription soon made this OCR-B setup with external companies obsolete. The advent of general OCR program packages for the PC later on made OCR topical again. The affordable software Omnipage let us convert printed text with diverse fonts to machine readable files. The Norwegian Institute of Local History has compiled a list of printed historical sources containing some 2,000 entries, useful when selecting material for optical character recognition. A few pilot projects used OCR to computerize printed source copies, but nowadays it is more efficient to let local historians and other volunteers transcribe source materials with their PCs. A noted example is from the National Archives' collection of medieval documents, with the 19th century publication *Diplomatarium Norvegicum*. Indexes to the collection were constructed by transferring the printed versions to text files with OCR. This increased the typical number of documents used for a dissertation from a few to a hundred. Another significant collection of digital source material only mentioned here even if it may contain microdata, is the collection of computerized records retrieved from public agencies by the National Archives and preserved in secure systems there (Thorvaldsen, 1992).

3.3 STANDARDIZING AND ENCODING THE CENSUSES

The transcribed censuses exist in a verbatim full-text version as well as in an encoded standardized format. It contains numeric codes for occupation, family status, and parish or municipality of birth, and no numeric codes were transcribed from the sources. In first instance, the main purpose of the coding of the censuses was to create statistics. There are several reasons for not simply using the aggregates published after each census. Boundaries between the administrative census units as well as the categorizing of occupations and other information between the censuses often changed, making historians' comparisons over time difficult. A third reason is that the variables in published statistics are combined at group levels, not at the individual level, increasing the risk of introducing ecological fallacies, i.e. jumping to conclusions based on aggregates (Langholm, 1976). A fourth reason is that the standardized codes make it easier to link people from source to source, for example code 0724 for the birthplace is more consistent over time than changing municipality names for the same locality.

While simple fields such as gender, marital status and age require little standardization and few rules, coding occupations, family status, places of birth and ethnicity are complex tasks (Thorvaldsen, 1994). After verbatim transcription, we use computer programs to semi-automate the coding. By eliminating identical versions of the same occupation etc., thus compressing the source entries of each variable into frequency lists, the coding became more consistent and less time-consuming. Each person entry is equipped with relevant codes, creating a standardized version of the census that can be used on its own in a statistics program or together with the verbatim text version of the source for record linkage.

Relationships between people in the same household or family are coded with specially constructed variables by the IPUMS project after they received our abovementioned encoded versions. For example, a relationship variable provides information about the spouse because it reciprocally contains the ID numbers of the spouses of married persons in each household. Using these ID numbers makes it clear which husbands and wives belong together, even if there were several couples in a household. Corresponding variables "point" from the children to each of the parents. However, in large, complicated households, the IPUMS computer program that creates these variables may introduce errors, which can be corrected at the Histreg.no website. The relationships between household members in the census lists allowed us to distinguish between the 19th century decline in the number of farm servants, and the domestic servants whose numbers did not decline until World War II (Thorvaldsen, 2008).

"Place of birth" is an important variable both because it can distinguish between people with common names, and it helps us to get an overview of the life course of migrants. A Norwegian census will usually specify a person's place of birth at the municipality or parish level. All municipalities in the same county have the first two digits of the four-digit code in common, which simplifies the study of migration. When the third digit is zero, it means that the relevant municipality is urban. Thus, historically there were never more than 9 towns or 99 municipalities in a province. To track administrative border changes, the numbering system is dynamic. Each area that has ever been a separate municipality has been assigned a unique code. There are lots of municipality changes, while the counties seldom changed. To handle immigrants, the system has been expanded with country codes for the rest of the world.² Problems can arise with ambiguous names of municipalities and other place names, as we shall see below. Sometimes, the name of a territory that includes several municipalities is indicated. If the entire area lies within the same county, the county code can be used.

Table 1 presents an overview of the encoded data from the censuses in 1801, 1865, 1875, 1900 and 1910 and hopefully soon 1920 as they are available via the internet from our partner at the University of Minnesota (ipums.org and nappdata.org).

Table 1 *Coded text strings and population figures in the national censuses 1865 to 1910 and Sandefjord town in 1920. The full count resident population in 1920 was 2,649,775*

Census variable	1865	1875	1900	1910	1920
Family/household	19,188	22,000	25,172	20,050	1,390
Occupations	75,734	78,000	368,598	334,079	2,044
Birthplaces	41,813	55,000	73,893	58,227	1,083
Population	1,701,756	1,813,424	2,240,092	2,391,782	5,764

Standardization of name strings is useful for quantitative analyses of name frequencies as well as for linking data records. Both first and last names have been standardized, in collaboration with professor of Nordic languages Gulbrand Alhaug (2011), in a research council-funded project. This happens via a model with three levels:

A. Graphemic level <i>Orthographic variants</i>	B. Phonemic level <i>Linguistic variants</i>	C. Lexicographic level <i>Standardized name</i>
Caroles Carolus Charolus	Karoles Karolus	Karolus

The project prepared a list of personal names where variant spellings and nearby linguistic variants were standardized to the same standardized name. For example, Kristian with K and Ch was coded to the name Kristian, and Fredrek is standardized to Fredrik, while the difference between Anne and Anna was preserved since the variants are pronounced differently and both variants are common. Since the standardization is rather conservative, it is necessary to use an additional program that calculates the phonetic distance between name forms, using the Jaro-Winkler algorithm (Winkler, 1990). Thus, personal records with similar name forms will be linked even though the initial standardization was conservative. A comparison of the effect of name standardization on linking in the US and Norwegian censuses showed that the number of links in the Norwegian ones increased by 17% due to the standardization and that it was the work with surnames that contributed the most (Vick & Huynh, 2011).

2 For Norwegian and foreign historical place codes, cf. https://rhd.uit.no/koding/fs_koder.html.

Another factor is that naming customs change over time (Fure, 1990). This affects both the frequency of different names in different parishes, and which forms of name are perceived as synonymous.

"Farm data" found in the 1838 and 1886 tax lists were transcribed with OCR or manually and are available on the Internet. The census of 1865 and 1875 also include agricultural data. Compared with the tax lists they are more complete, including cottars *and* farmers and also provide information on the seeding and the number of animals. However, the reliability of these figures is weakened because the informants feared taxation. So, one may assume that the census data on sowing and animal husbandry give a deflated picture. But it is still realistic to construct a measure for relative production at the farms and cottars' places. In principle, we considered two calculation methods, either the monetary value of the production or its nutritional content. Because there was no real market for potatoes, grain etc. in much of Norway, we based the calculations on the nutritional value of which the number of calories was the most important aspect (Statistics Norway, 1880). These were calculated and considered in detail in the research to assess the value of the introduction of potatoes in Norwegian agriculture (Lunden, 1975).

For husbandry we use the calculation method from Statistics Norway for the 1875 census. The conversion of the relative value of different livestock into cow-entities is based on sales value for adult animals and thus takes account of local differences (Thorvaldsen, 1995b, pp. 486–488). Again, we use Lunden's (1975) calculations to estimate the farms' and cottars places' production value as thousands of calories per year. In the 1865 census, the agricultural figures were entered in the same form as the rest of the information and are thus linked to individuals, mostly heads of household. This is more complicated in the 1875 census because agricultural data was noted on a separate form but can be linked automatically to persons and places. For instance, a calculation of agricultural output was performed in the community history of Kvenangen municipality to assess differentiations in production connected to the three ethnic groups in the area, Sami, Fins and Norwegians (Bjørklund, 1985).

4 THE HISTORICAL POPULATION REGISTER (HPR)

4.1 THE REGISTER

The HPR is becoming a national register covering the 9.5 million people, who lived in Norway sometimes during the period 1801–1964 with an estimated 87 million person records in the most important cross-sectional, vital events and migration sources (Holden, Boudko, & Thorvaldsen, 2020). The population register has an open access part including only deceased persons in open sources. Except for the deceased from 1928 to 2014, there are few open sources after 1920. The closed part contains restricted sources and persons still living. Both parts are linked to the Central Population Register (CPR) started in 1964. Automatic record linkage has been made at the UiT The Arctic University of Norway and at the Norwegian Computing Center (NR). NR has also developed the website histreg.no, where manual links in the open part are added by volunteers. The links between the open and the closed parts are not public, however. Most censuses and church records from 1800 to 1960 have been transcribed and the open parts are becoming available in the Digital Archive (Holden et al., 2020; Thorvaldsen, 2011b).

The motivation for building the HPR is to provide a central national infrastructure for research in history, social sciences, medicine and a number of other disciplines. The HPR also has an important cultural component of cooperation about the tracing of genealogies for over 200 years, and by comparing inconsistent records it fulfills a crucial source critical goal. In Section 6 we will present methods by which automatic links are created between instances of the same person in different sources and pointers to relationships between family members. This is done both by UiT and NR. In addition to the above-mentioned institutions, the National Archives (the Digital Archive), the Institute of Public Health (digitizing and linking 20th century sources), Statistics Norway (access to the Central Population Register), the National Library (support of and interaction with local historians) and the Norwegian School of Economics (tax records) are project partners. The Historical Population Register recently received its second significant funding from the infrastructure program of the Norwegian Research Council.

Linking of instances of persons in various sources has traditionally been carried out in the context of farm and family history for rural municipalities with detailed studies of the local church registers and censuses supplemented with other written and oral sources. The studies have mapped settled farmers to a greater extent than people without real estate and people who moved. In most cases,

such work is carried out by people with detailed insight into the local conditions. Traditionally, the work has been carried out manually, but computers have increasingly been used to streamline and systematize the work with the source material and the analyses. Machine representation also simplifies the communication of the results (Kjelland, 2018), usually in community history books. It has long been a goal among professional historians to activate the community history books' detailed farm and family genealogies in historical research (Hovland, 1977), and we believe that the cooperation with the National Library will promote this aim.

4.2 OPEN ACCESS TO THE HISTORICAL POPULATION REGISTER: HISTREG.NO

The open part of the Norwegian Historical Population register is available at the website histreg.no. Introduced in 2016, the National Archive is responsible for the site, which is developed by the Norwegian Computing Center. Histreg.no may be considered as an index to the Digital Archive of the National Archive, with transcriptions of church books, censuses and emigrant lists with about 57 million person-records from the period 1800–1920, as well as sources like prison and health records, school protocols, etc. Histreg.no has a page for each person entry in all these sources. If several person entries are considered to belong to the same person after record linkage, the pages are merged so that the person page presents the life course, including a list of all sources belonging to the person. Each person gets a unique ID generated from the unique source entry ID provided by the National Archive. The use of this unique ID in scientific articles and elsewhere enhances the documentation of the data used in research. The URL to the person page of the explorer Fridtjof Nansen is <https://histreg.no/index.php/person/pf01073681015788> where the last 16 digits are his unique ID which in this case originates from the 1920 census in the Digital Archive. For each source entry, there is a hyperlink to the transcribed source in the Digital Archive and the scanned source image.

Figure 2 shows a typical person page in histreg.no, the top showing gender, name and information about the birth and death. This information may be edited by a logged in contributor. Then follows links to the person pages of parents, siblings, partners and children with information about family relations from the sources and additional family relations added manually. The life course table lists the linked source records about the person with hyperlinks to the transcribed sources in the Digital Archive. This information may only be changed by adding or removing source records, not by changing the data in the sources. At the bottom of the page (not shown in Figure 2), it is possible to write a brief biography, explain the linking, specify references or add other comments about the person.

The record linkage algorithms to create unique persons from several appearances in the sources are developed by the UiT The Arctic University of Norway and the Norwegian Computing Center, where the record linkage rate varies by source. It is close to 90% between the 1910 and 1920 censuses where we have families and birthdates and significantly lower for sources with less information. The algorithms are based on comparing the names, birthdate or -year, birthplace, address and family relations (see Section 6). In addition, volunteers make links and family relations manually in a crowd sourcing effort. During recent years, more than 170 persons made about 40,000 links per month corresponding to the output of two persons working full time in the same period. Still, more than 90% of links are made by algorithms. By March 2023, the Histreg database contains 12,2 million links between source entries for 3,7 million persons (i.e., persons found in at least two sources). We estimate there were 6,6 million persons living in Norway in the period 1800–1920 (Statistics Norway, 1995). Each manual link and family relation has a time stamp and identification of the contributor. The same person is sometimes registered with conflicting data in the sources. The system, therefore, includes the option to register a link as verified in spite of conflicting information, in order to block the link from accidental removal during quality controls. Histreg also lists conflicting data that are not verified, which manual contributors are encouraged to check manually. We expect to continue adding millions of links during the following years, but Histreg will never be "complete".

Histreg.no has many features to improve the quality of the dataset and encourage its use. It is possible to refer to persons that are not yet identified in the sources. The program may list the largest unlinked families in a specific census for a municipality in order to encourage further linking in a region. In addition to the sources including the whole population we include some thematic registers, such as war prisoners 1940–1945. There are also links to the Local History Wikipedia and biographic data in newspapers.³ During 2023, we shall add information on wartime sailors and Norwegian politicians

3 https://lokalhistoriewiki.no/wiki/Lokalhistoriewiki:Hovedside/Om_Lokalhistoriewiki and <https://www.retrievergroup.com/about-us>

from 1814 onwards provided the legal restrictions are heeded. Complementary sources provide further information about the persons, enhance the value of each thematic register and increase the interest in Histreg since mutual links make the thematic register more visible. At the same time, the protection of privacy must be respected for persons still alive.

Persons linked in Histreg are not representative of the entire population, as in all historical population registers, since linkage rates are higher for persons with relatively high social status, a permanent address or (slightly) being male. However, the representativeness of statistical results from Histreg can be increased based on data from the full count censuses. Histreg is both an editing and a retrieval tool. In addition to personal information fields such as name and age, the user can refine the search based on type of event, role and year or period and geography — municipality of birth or residence. It is possible to show what partners and parents were related to the retrieved persons. Histreg.no can sort the search results by first name, surname and year of birth. The search also shows the number of interlinked person records.

Figure 2 An editable person page in Histreg with keywords in English

Ida Mathilde Mikkelsen Hansdater

♀ Kvinne
 Født: 08.07.1886, Sem sogn Born
 Død: 06.04.1962, Ukjent sted Dead

[Rediger persondata](#) [Søk etter lignende personer](#) [Edit person page](#) [Search for similar persons](#)

Foreldre og søsken Parents and siblings

Far: Hans Jörgen Kristensen Christensen , 1853 -
 Mor: Anette Andrine Regina Kristensen Kristensen* ChristensDater , 1861 -
 Søsken: Kristian Hansen Hansen
 Søsken: Agnes Hansen
 Søsken: Anna Marie Hansdtr.
 Søsken: Hans Ragnvald Hanssen , 1888 -
 Søsken: Johan Arne?? Hanssen , 1893 -
 Søsken: Wilhelm Kristensen* Wilhelmsen , 1894 - 12--
 Søsken: Hilda Anette Hansen , 1896 -
 Søsken: Georg Kristensen* , 1896 -

Partner og barn Partner and children

Partner: Gjert Mikael Mikkelsen , 1884 -
 Barn: Helmer Thorvaldsen , 1912 -
 Barn: Gulborg Thorvaldsen , 1914 -
 Barn: Maud Ida Thorvaldsen , 1918 -

[Vis lenking av familiemedlemmer](#) [Show linking of family members](#)

Life course table - Overview of sources with this person

Livsløpstabell - Oversikt over kilder med personen

Nr	Dato	Kilde	PFID	Rolle	Navn	Fødselsdato	Fødested	Bosted	Fam. stil.	Sivst.	Yrke
1	01.01.1891	FOLK, Sem, 1891-1891	pf01052816001475 Census		Ida Mathilde Hansdater	1886	Sem	Sem: Langerød (Husmondsp)	Datter		
2	03.12.1900	FOLK, Sandar, 1900-1900	pf01037148003376 Census		Ida Matilde Hansen	1886	Sem	Sandar herred: Lystad	d	ug	Datter
3	01.12.1910	FOLK, Sandefjord, 1910-1910	pf01036489005150 Census		Ida Mathilde Mikkelsen	08.07.1886	Sem sogn	Sandefjord: Kongens gate 37	hm	g	Hustru
4	01.12.1920	FOLK, Sandefjord, 1920-1920	pf01073803003878 Census		Ida Mathilde Thorvaldsen	08.07.1886	Langerød i Sem Vest	Sandefjord: Torvgaten	hu	g	Husmor
5	06.04.1962	DODR, 0000, 1951-2014	pc00000001724177 Church record	avdød	Ida Mathilde Thorvaldsen	08.07.1886					

4.3 OTHER USER INTERFACES

4.3.1 LINKED PAIRS OF CENSUS RECORDS

The simplest way to follow people over time is to link two points in the life course, for example a baptism and a census or two censuses. The latter is done for the censuses of 1865, 1875, 1900, 1910 and 1920. The first three of these enumerations were linked and made available by the Minnesota Population Center as part of the North Atlantic Population Project (NAPP). The website ipums.org contains data files for Norway with links between the censuses 1865 and 1875, 1865 and 1900 as well as 1875 and 1900, which have been imported into the Historical Population Register. In addition, NAPP includes linked records combining the complete US 1880 census with seven US census samples as well as Norwegian census records (cf. https://international.ipums.org/international/linked_data.shtml). The Norwegian and American linked censuses have been used together in research on the economy of emigration (Abramitzky, Boustan, & Eriksson, 2012, 2013). On nappdata.org, the linked censuses can be downloaded in a simple data format i.e., using one record for two linked data records; this simple data structure is possible when only two points in time are covered in each life cycle.

In order to avoid constructing erroneous biographies by linking records that actually belonged to different people, a conservative linking strategy was chosen by the NAPP project, which resulted in low linking rates (Ruggles, Fitch, & Roberts, 2018). The linking strategy for the Norwegian censuses 1865–1875, 1875–1900 and 1865–1900 depends on four time-invariant variables: year of birth, four-digit municipality birthplace code, standardized first name and standardized surname. Birth years were allowed to differ by up to three years for linking men, and up to five years for married couples. Some municipal border changes were neutralized by including the neighboring municipality. To avoid creating a biased selection by under-prioritizing the linking of singles, information on family members was not used – except when linking married couples. Unfortunately, single women were more difficult to link, due to the lack of birthdates in these censuses (Thorvaldsen, 2011b). If the same data record was linked to two different records in the 1875 or 1900 censuses, because these censuses combined a de facto and a de jure count of both resident and present population), information on permanent residents was preferred (Thorvaldsen, 2006). The early Norwegian immigrants to the US, coming before the keeping of systematic migration protocols, have been listed. We have also successfully traced emigrants to Sweden and to north-western Russia (Naeseth & Hedberg, 1993–2008; Thorvaldsen, 2011a; Thorvaldsen & Erikstad, 2007).

4.3.2 TIMELINES TO FOLLOW INDIVIDUALS AND FAMILIES

The UiT The Arctic University of Norway presents a system to display "timelines" with longitudinal information from the Historical Population Register (Thorvaldsen, Sommerseth & Holden, 2020). The 1865, 1875, 1900, 1910 and soon the 1920 censuses are available for search on the UiT website via a simple and an advanced user interface, see <http://rhd.uit.no>. After finding a person in one of the enumerations, clicking the house symbol displays information about the household. Linked individuals are equipped with a marker (🏠), shown in the left margin in Figure 3, meaning that further information is available from other sources via unique source references generated by the National Archives.





By clicking the link marker, an overview of the linked data records will be displayed from other censuses and church records. Users are warned that the links are generated automatically, and it cannot be ruled out that the software has introduced erroneous links and that some are missing. For example, Thorvald Mikkelsen was easily identified in the censuses in both 1865, 1900 and 1910, but was not *automatically* linked to his entry in the 1875 census because no close relative was present to be used as linkage criteria. Since he remained on the same farm, the linking was easily done manually. He was adopted as a foster child by the childless farmer who bought the farm from his parents. Clicking on the + sign in the column on the left in Figure 4 displays information about the entire household to which the person belonged in the relevant census year.

The timeline function makes it less time-consuming to follow groups of people over time. Cohorts of people who share the same characteristics can be defined in the advanced user interface, so that the search results are more adapted to statistical purposes. However, there are no built-in statistical procedures, and the user must create the categories herself. An example of a more complex timeline, which also contains information from the church records about fisherman Haldor Hansen (1850–1922) can be found at <https://rhd.uit.no/folketellinger/tidslinje.aspx?idi=8231280>.

Figure 3 *Family on the farm Sjuvestok in Stokke parish in 1865 with linkage symbols in the left margin*

Name	Family status	Marital status	Occupation	Birth year	Place of birth
 Mikkel Nilsen	hf	g	Gaardbr. og Selveier	1812	Stokke Pr.
 Elen L. Hansdatter	Hans Kone	g		1823	Stokke Pr.
 Hans Mikkelsen	Deres Søn	ug	Matros	1844	Stokke Pr.
 Mathias Mikkelsen	Deres Søn	ug	Hjelper Fdr. med Grbr.	1847	Stokke Pr.
 Thorvald Mikkelsen	Deres Søn	ug		1849	Stokke Pr.
 Rikard Mikkelsen	Deres Søn			1852	Stokke Pr.

Figure 4 *Timeline with references to four censuses for Thorvald Mikkelsen. It can be expanded for each census at <https://rhd.uit.no/folketellinger/tidslinje.aspx?idi=3177832>*

	1865	FT1865	0720-001-0027-00-005
	1875	FT1875	0720-001-0103-00-003
	1900	FT1900	0718-002-0075-00-001
	1910	FT1910	0706-005-0038-05-001

For advanced statistical purposes, the advice is to use the Intermediate Data Structure (IDS), which was specified by researchers and data providers who need to transfer records between collaborators (Alter, 2021; Quaranta, 2021). For qualitative purposes the Linked Pair approach or the Time Lines described above, provide a simpler introduction to using the HPR. The IDS has been successfully implemented for regional data from Northern Norway at the Norwegian Historical Data Centre and used to find ground-breaking results on intergenerational infant mortality, in an international project also studying regions in Sweden, Belgium and the Netherlands with comparable datasets (Quaranta & Sommerseth, 2018; Sommerseth, 2018). This is not the place to describe the qualities of the IDS system, which is well documented elsewhere (Alter & Mandemakers, 2014). However, IDS is not a standard tool in Norwegian research on historical microdata. The development of alternative models for data exchange between the partners may serve the internal project needs and may also be a way to distribute microdata to researchers in Norway. But this will certainly not function as well as IDS to promote internationally comparative research projects. IDS is designed for family reconstitution data and has less advantage for census data. As the Historical Population Register now adds relatively more information from vital registers, IDS will likely prove more useful.

5 GIS AND MAPPING

5.1 OVERVIEW

The starting point for work with GIS in connection with microdata in Norway was the Municipality Database (*Kommunedatabasen*) built and maintained by the Norwegian Centre for Research Data containing statistical information about the municipalities since the 18th century.⁴ This database was exploited successfully by the Princeton Project-related fertility decline study, at the time when full count microdata for Norway was still not complete (Sogner, Fure, & Randsborg, 1984; Sogner, Randsborg, Fure, & Walloe, 1986). Attached to the database is a dynamic collection of national maps showing the province and municipality borders dynamically over time since their creation in 1837, which can be ordered and downloaded for a number of GIS software platforms. The UiT extended the municipality map backwards to 1801 with the parish boundary maps from the transcribed 1801 census (Statistics Norway, 1980). Besides, the National Archives, the Institute of Local History and others

4 <https://www.nsd.no/finn-data/kommunedatabasen/>

have cooperated to develop a dynamic cadaster of farms and their placement in the administrative boundary structure since the 18th century, but this project has not yet come to fruition.

A considerable part of Norwegian historical research deals with local areas, municipalities, provinces and other regions below the national level. Therefore, it is important to pose some questions regarding geographical delimitations: How shall the area of study be defined in relation to the surrounding area? What criteria might be used to partition the selected area internally into smaller zones for more detailed study? These questions are closely related to the use of GIS software to display characteristics of population differentials on choropleth maps. The divisions into geographic entities were made through a combination of political, administrative and juridical decisions. More recently infrastructure, social and cultural divisions play a role as well, although language and ethnicity are seldom used as border criteria in Norway.

Both because of and in spite of the rather drastic alterations in municipality boundaries, the four volume *Tromsø City History* is based on its present-day borders. Pedagogically, most readers are more familiar with current municipality boundaries and the financing of the project would have been difficult if the historical work covers areas outside the present municipality or excludes part of it. In volume II Astri Andresen (1994) used the census microdata from 1801, 1865, 1875 and 1900 to reconstruct the present-day boundaries of the municipality for the long 19th century. She selected the data for the specific farms or places situated in the area that today is a part of Tromsø. These contiguous pieces then became censuses covering the present-day Tromsø municipality over time. In contrast to the printed aggregates from Statistics Norway, where the old boundaries were followed, she made overviews of parts of the area or all of it. Thus, we get to know how many people lived in today's Tromsø-area on past census days, the birthplaces of the inhabitants according to 19th century censuses, etc. The ethnic composition of the population was studied by analysing the Sami areas in peripheral parts of the municipality. There are valid reasons why one should use the present-day administrative divisions by back-projecting them to earlier periods: Access to the sources might be easier, the provenance principle decreeing that source material is to be organised by the present-day administrative divisions. Thus, a discussion of the choice of region is imperative, but is often lacking in commissioned research.

The reduction in the number of Norwegian municipalities by nearly 50% of their maximum number, confirms that boundaries on a low level are not stable over time. The boundaries of the provinces were traditionally more stable than the borders of the municipalities, but they too have been subject to small changes, sometimes when the municipality-boundaries have been changed. However, this changed dramatically when a conservative Parliament in 2017 supported a regional merger reform act that affected 13 out of 20 provinces, an unpopular measure in most provinces. After new elections, the centrist parties dominated Parliament and voted to dissolve most of the mergers. Future historians will, therefore, face extra hassles when mapping developments on the provincial level around 2020.

The complicated process of determining the borders between administrative units started in medieval times and will not be detailed here. It must suffice to start with the Laws of Local Democracy (*Formannskapslovene*) of 1837, which mainly based the new municipalities on the old ecclesiastical divisions into parishes and sub parishes (*prestegjeld* and *sogn*). At the time, the border surveys were provisional due to lack of resources to stake the frontiers with locals in the field or examine the relevant archives. The established boundaries only gradually became more certain, as borders were regularly revised based on new information. A game changer was the detailed descriptions of the provinces in *The Country and People of Norway* (Helland, 1876–1917).

5.2 SPLITTING MUNICIPALITIES FOR ANALYSIS

In connection with local micro-studies and as a basis for comparison, it is necessary to divide the area of study into lesser units. Here the commissioned researcher will have more freedom. When writing his volume III of the history of Oslo, *The Divided City*, Knut Kjeldstadli (1987) discusses several ways of dividing Norway's capital, singling out five factors that to varying degrees have contributed to the actual formation of the city's districts: boundaries, names, local institutions, administrative divisions and social conditions.

A source-oriented approach may work to some degree for rural municipalities. In the *History of Balsfjord and Malangen Municipality* Hauglid's (1981) attempt to characterize the social history of each particular census ward on the basis of the machine-readable versions of the nominative censuses of 1865 and 1900. In the chapter called "Ethnic Diversity in a Multi-Cultural Society" the census of 1865 is the main source along with Friis' ethnographic maps from 1861.⁵ The author examined each of

5 <https://www.dokpro.uio.no/friiskartene/1861/1861oversikt.html>

the four census tracts to map settlement patterns, the distribution of Norwegians, Sami and ethnic Fins and their kinship relations. Aggregative figures were directly derived from the census, although with explanations based on other source materials. Because of changed boundaries this 1860s snapshot is not comparable with the one from 1900 with 16 census wards. This unfortunately urged the author not to aim for consistency and comparability over time, but rather highlighted occupations than ethnicity. Thus, even though we get much insight into local ethnicity as well as trades and industry, a systematic description of industrial and ethnic developments is regrettably lacking.

In *The History of Sandefjord* Finn Olstad (1995) took this one step further by attempting to divide Sandar municipality, which used to surround that town, into a coastal and a land-locked part. His purpose was to investigate any differences in trades and industries in separate parts of the municipality according to the 1900 census. About half were employed in the primary sector, mainly agriculture in the land-locked part of the municipality, while this was only true for about a fourth of the population in the coastal part, rather partaking more in maritime trades. However, this was based on the census wards which often do not follow such a divide, but rather extend from the coast into the inland areas. An alternative strategy has been to rather use the railway line and the main road as demarcation lines parallel with the coast, but this necessitates the use of farms and other place entities in the census to divide the municipality into a coastal, an internal and a middle zone between the railway and the main road. Such a definition is easy to relate to for most readers (Thorvaldsen, 1997).

With its Ward Database (*Kretsdatabanken*) Statistics Norway and The Norwegian Centre for Research Data made information on the level of census wards more accessible for researchers, but only for the post-war censuses. The Wards Database mirrors the official statistics from the censuses on the municipality level: the population's gender and age structure, industry and employment, religion and language, as well as housing with up to 500 variables. Thus, the census wards which only had a practical function for the census takers, became statistical units of analysis. As part of this, Statistics Norway redesigned the boundaries of the census wards in order to create meaningful entities, e.g., by distinguishing between urban and rural areas. This could render aggregates from the wards less comparable over time unless researchers re-aggregated the information according to the new boundaries. For foreign researchers census ward data is an interesting alternative, since the Nordic countries do not participate in the IPUMS.org project with microdata from the post-war period. However, the online database microdata.no where users can create anonymized aggregates, is not available to researchers without a Norwegian social security id-number (Ballo, 2019). The relative inaccessibility of the latter dataset does not warrant a detailed description here.

5.3 CHRONOLOGICAL ADJUSTMENT OF THE MUNICIPAL AGGREGATES OR BOUNDARIES

In national analyses, social scientists and historians have often employed the municipality as a unit for geographical data. This is because the municipal level is seen as less random and more pedagogical than other divisions, and large amounts of statistics and other information are available on the municipal level. However, the constant changing of municipal borders must be solved if we wish to use these data sets for comparison over time. A standard solution is integrated in the Municipality Database of the Norwegian Centre for Research Data, containing variables about Norwegian municipalities from 1769 onwards. When extracting a time series, researchers can request that the municipal boundaries are to be standardized to a chosen year. The software will then "move" a proportion of the population affected by the border changes in the period studied. This method can be illustrated with a simple example from election studies, where we want to study the development of voting from 1933 to 1936. This is complicated because an area with 100 persons was transferred from the municipality Fjord to Fjell in 1935. If we do not have data for the transferred area, we must assume that the votes from the group of 100 were distributed in the same way as the votes in the entire municipality. If party A received 70% and party H received 30% in Fjord in 1933, the program moves 70 A-votes and 30 H-votes from Fjord to Fjell in 1933 before comparing with the 1936 results, obviously a gross approximation with potentially misleading results. Often, the transferred areas are peripheral rather than central parts of the municipality and have a different employment structure. The splitting of a municipality into two new ones can be similarly difficult to approximate, whereas the merging of such administrative areas is more straightforward.

Ideally, we should study migration and other social phenomena in relation to the smallest locations, that is, to the farm or the building (Thorvaldsen, 1995a), which is difficult, since the census only reports place of birth on the municipal level. The study of migration in Troms province from 1865 to 1900 was complicated

by the changes in the parish and municipality boundaries. Thus, an important prerequisite when studying migration, is to create a consistent division of the province into municipalities over time. The census of 1865 was taken for thirteen parishes in Troms province, while later censuses used a more fine-grained division. For this reason the rough parish division of 1865 was used as a basic structure to make comparisons over time. By "moving" the farms and other places into the 1865 parish structure, it was possible to merge the smaller municipalities in later censuses into a common structure for the whole period.

When we transfer people from one municipality to another to compensate for changes in the municipality borders, we must be certain to change people's place of birth accordingly.

5.4 THE MAP PORTALS URBGIS AND BERGIS

The UrbGIS map portal shows historical maps for the urban areas in Norway and is integrated with the 1900 and 1910 censuses in the Digital Archive. The map for Bergen in UrbGIS mainly covers the period 1881–1957 based on street names and numbers. The richer collection of historical maps for the city of Bergen in BerGIS covers the period approximately 1830–1881 using both the address system and a cadastral system. The map portals can also show map layers for applicable property boundaries (see <http://urbgis.uib.no/> and <http://urbgis.uib.no/bergis/>).

5.5 ENUMERATING THOSE ABSENT AND VISITING

In the censuses from 1875 onwards, some persons according to the instructions were enumerated twice. These were people travelling or away from home for other reasons, most often to work in a different location or to live preliminarily away from home. This has consequences for our statistical use of the census microdata. Since no one should be counted twice, we must choose between the two entries. And a similar decision must be made during record linkage as discussed in Section 6.

Earlier, in the 1865 full count census, census takers were instructed to enumerate residents "where they sleep" and not include "Anyone residing temporarily in a place, ...". Thus, this was a *de jure* census, unlike the British at the time, which noted people where they happened to be on census night, i.e. *de facto*. Problems, especially with the registration of sailors, may explain why combined *de jure* and *de facto* enumeration was introduced into the 1875 census. Inspired by the international statistical conferences, the 1875 census forms contained a special field for visitors in order to note the "usual residence of those, who on the 31st of December temporarily stayed overnight in the house." For those temporarily absent, a special section was included at the form's bottom, an arrangement dropped in later censuses. Statistics Norway detailed the definitions of temporary residence or absence, e.g., about absent lodgers. Students and servants in 1875 should be enumerated *de jure* where they studied or worked. In later censuses, however, the students were considered permanent residents in their municipality of origin, temporarily absent from there and temporarily present where they studied. As their numbers increased after 1960, the university cities successfully lobbied for a change in the enumeration rules in order to receive state funding according to population size. Thus, for the 2001 census, statistics Norway again enumerated students where they lived when studying (Thorvaldsen, 2006).

When analysing the censuses from 1875 onwards with microdata, it is important to choose whether to exclude either the persons absent or those visiting. Including both leads to over-enumeration because of duplicates. Including those absent results in the resident or *de jure* population, while including those visiting gives the present or *de facto* population. In-between groups such as the abovementioned students can be fitted into either category according to national specifications. Especially in municipalities with many sailors or fishermen, the difference in population size according to the definition of the resident and the present populations can be significant, even exceeding a tenth of the population. It follows that men predominated (see Table 2). It is possible to perform record linkage between records about those visiting and those absent in the same census. Complete matching of the groups cannot be expected, since many of those absent had left Norway and there are foreigners among those visiting. Also, inaccuracies produce the usual problems linking nominative microdata. If we demand record linkage matches where the years of birth and birthplace codes are identical, and a Jaro-Winkler similarity between first names in the records for the visiting and absent groups, these are matched in 29,618 cases. If we demand exact match on birthdate, the number decreases to 20,210 — it is natural that there are more problems with birthdates for these mobile groups than among those who were enumerated at home. Further analysis of these records should be undertaken in order to assess the proportion of inconsistencies between enumerations by different census takers at the same point in time.

Table 2 *Number of persons by gender and residential status in the 1910 census for Norway*

de facto/de jure	Sum	Female	Male
At home	2,256,281	1,205,084	1,051,183
Visiting	133,591	29,708	103,881
Absent	82,222	25,902	56,317
Sum	2,472,094	1,260,694	1,211,381

6 RECORD LINKAGE SUMMARY

Many articles and several books have been written about the linking of historical individual data. We covered basic points above in connection with the Historical Population Register. It is an impossible task to discuss all the rules and experiences here, but based on our practice the main principles can be summarized in ten points (Fure, 2000; Thorvaldsen et al., 2015).

6.1 TEN COMMANDMENTS OF LINKING HISTORICAL PERSONAL RECORDS

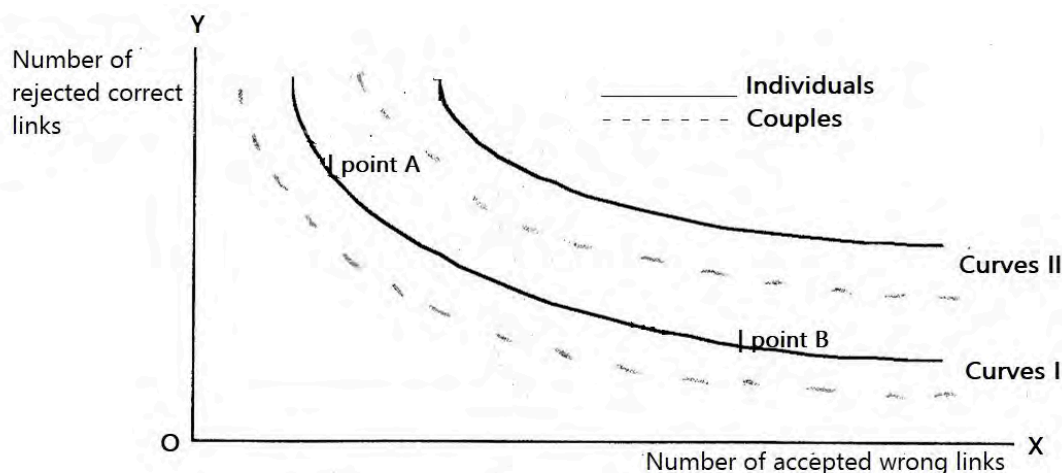
1. The life course that the links describe must be logical. Obviously, a marriage record must precede a burial record.
2. Historical, individual level microdata contains too many inconsistencies to demand full agreement between the variables used for linking records from two or more sources. For example, there will often be variant spellings of names or places of birth and errors in the year of birth.
3. Only accept reasonable discrepancies between the source entries based on experience, e.g., with name variants. Kristine and Kristina should be considered true name variants, but Anna and Anne are both frequent and distinct.
4. We cannot link duplicate identities, only classify them as linking candidates. An example is several persons named Ole Olsen born in Oslo in 1851.
5. Search for complementary source material in case of doubt (Point 3 and 4). For example, birthdates in the 1910 and 1920 censuses could provide a basis for linking back to a baptism list with birthdates that could not be linked to the 1900 census.
6. Variables that vary throughout the life course may be used source critically for linking, the most relevant time-variant information being addresses and occupations. We can therefore link an apprentice carpenter's record to a later entry with a master carpenter even if there are other competing entries with different occupations but otherwise duplicate information.
7. Relationships to other people may also be time-variant, but can still be used critically for linking, even if information on group relations is not constant over time in the sources. Links made with time-variant variables should be flagged (cf. the end of Section 4.2).
8. The links can be changed manually, usually based on information in genealogists' records. However, this must be flagged and documented, and it may have consequences for other links in the database (cf. Point 10). Links made according to Rules 6, 7, 8 may create bias in the linked sample of records.
9. The precision of protocol data generally improved over time. Thus, you can place greater emphasis on small differences in names, year of birth, etc. when linking records with similar information around the year 1900 than around the year 1800 and still avoid the risk of linking duplicates (cf. Figure 5).
10. Some links will be broken because new sources are added, algorithms are improved or genealogies checked. It can potentially lead to many changes — like a nuclear proliferation in the database but will usually have smaller consequences. Even so, the consistency of the links in the database must be periodically checked with detailed algorithms.

Below we shall deal with typical examples of problems and solutions that arise when linking the 1910 and 1920 censuses into the Historical Population Register. These censuses are the first that contain exact dates of birth. This provides unique opportunities for analysing discrepancies between different sources with personal data. When linking retrospectively, the 1900 and 1891 censuses as well as the church records can be brought in to resolve discrepancies. In this way we fulfil parts of the source-critical purpose, which together with the more empirical one, are the main motives for creating a historical population register.

Manual linking provides flexibility and makes it possible to consider special cases that are only described in the tradition of each individual family, local historical knowledge or special name forms of the nickname type. In addition to being more efficient, automatic record linkage using computer programs will ensure more uniform handling of the sources, but this can more easily introduce errors due to lack of consistency and uniqueness in the sources. Experience shows that the algorithms must be detailed in order to provide a constant proportion of correct links with time and place due to variations in the sources such as degree of name similarity, migration and other factors. When record linkage is based on the information in two or three source variables, these data items must be relatively unique and consistent with respect to name, date of birth, year of birth and place of birth as is often the case in such recent sources as the censuses from 1910 and 1920, and where we can check against supplementary sources and data on related persons.

The linking techniques are a compromise between creating as many correct links as possible while not introducing erroneous links, which is well illustrated for Denmark in Figure 5, a country with source material similar to Norway (Johansen, 2002). Here the relationship between the number of false positive (i.e., incorrect) links on the x-axis and the number of false negative (i.e., missing) links on the y-axis is illustrated. The main idea in Figure 5 is that stricter linking criteria will limit the number of false positives and increase the number of false negative links. Looser rules will have the opposite effects, increasing the number of false positive problem links, while limiting the number of false negatives. Johansen maintained that the increasingly exact content of the source material made it easier to find a compromise between the two considerations than was the case in earlier sources, illustrated by the two different sets of curves, marked I for the later and II for the earlier periods. The introduction of birth dates in the 1910 and later censuses was another step towards origin (O) in the diagram.

Figure 5 Linking individuals and couples



Note: Curves I: later periods, Curves II: early periods. Point A: Strict linkage rules, Point B: looser rules. Adopted from Johansen (2002).

6.2 RECORD LINKAGE PROBLEMS WITH THE 1910 AND 1920 CENSUSES

As stated above, the censuses of 1910 and 1920 were the first to ask for the date of birth for the entire population. The background for this addition was that from 1903 population registers were started in the municipalities, needing more precise identifiers. Name information in the censuses had become more unstable since marrying women changed their surname, and many persons abandoned the custom of patronymics and changed to other types of family names, for example the name of their farm. In addition, the number of persons born in cities increased rapidly during this period. The 1910 census

was transcribed in collaboration between the UiT The Arctic University of Norway and the National Archives. The combination of professional transcription staff and sources with good readability should guarantee a high quality of the work. The 1920 census was mainly transcribed by volunteers who were "paid" with insight into this source before the 100 years exclusion period expired on 3 December 2020. The relevant forms were scanned and posted in the Digital Archive with access for approved users.⁶ The quality of the scan is good, but the colour differences are gone, which makes it harder to distinguish between what is written during fieldwork and what has been added by local administrators or Statistics Norway. Altogether, there is reason to expect a somewhat lower transcription quality compared with 1910. It also plays a role that the 1920 census consists of one-person forms, where it is more difficult to decipher writing by comparing individual entries than when we have forms listing many persons, like in 1910. This can especially have an effect in town censuses because these were not filled in by rural teachers, but usually by the house owners themselves. However, we must bear in mind that a systematic overview of discrepancies in the 1900 and earlier census versions shows that most inconsistencies are not a result of transcription errors but of original source differences (Fure, 2000).

6.3 SURNAMENES, PLACES AND DATES OF BIRTH PROBLEMS IN THE 1920 AND 1910 CENSUSES FOR SANDEFJORD TOWN

6.3.1 DATE OF BIRTH

Of the 1,101 persons that were linked for Sandefjord between the 1910 and 1920 census based on name, place of birth code and year of birth, 180 had discrepancies in the date of birth. Although the "Date of birth" is a good distinguisher and helpful to find errors and inconsistencies in other information this variable is not always reliable either. We can further compare with the 1900 census which gave the date of birth for persons up to two years old. Trygve Gjertsen was born in 1900 in neighbouring Larvik town. His date of birth was the 1st of April in the 1920 census, but the 7th of April in both the 1910 and the 1900 census. A check against the 1920 census scanned image shows that a more likely transcription is the 7th of April. Eivind Halvorsen (PID pf01036489007059)⁷ born in 1900 in Sandefjord had two conflicting birthdates: 5th of June, 6th of June and again 6th of June in the three censuses 1900–1920. Since a question mark was added to the date in the first of these censuses, that is likely the wrong information. For Emil Larsen, born 1899 in Tjølling (PID pf01036489003252), the discrepancy is once more due to a transcription problem. As can be seen from Figure 6, the day digit '8' looks like '5'. Incidentally, Emil was called Hartvigsen after father Hartvig in the 1900 census, but in 1910 the whole family had changed their surname to Larsen.

Figure 6 Example of unclear birthday in the 1920 census for Emil Larsen

1 desember 1920.

Skjema 1. Personseddel nr. 3

Sandefjord by. Tellingskrets nr. 8

Husliste nr. 238 Husholdningsliste nr. 2

1. Fullt navn: Emil Larsen

2. Mannkjønn! Kvinnkjønn!

3. Fødselsdag 28-5 i året 1899

4. Fødested: Tjølling

Opgi herred eller by i Norge eller fødeland utenfor Norge.

6 <https://media.digitalarkivet.no/en/ft/browse?censuses%5B%5D=18&counties%5B%5D=07&municipalities%5B%5D=0706&text=>. The 1910 census is scanned and available in the Digital Archive, useful for control purposes.

7 We have added PIDs so that readers can study the examples by inserting them in Nansens url (cf. Section 4.2).

6.3.2 SURNAMENES

By 1920 it had become the norm to have a different surname than by birth, especially for women upon marriage. Ida Mathilde (PID pf01052816001475), born in Sem in 1886 experienced even four different surnames according to the censuses from 1891 to 1920. She was first entered with the traditional female patronymic Hansdatter in 1891, then with the male and more modern version Hansen in 1900. After getting married in 1910 the name became Mikkelsen, changed to Thorvaldsen in the 1920 census. Family tradition tells that the last change was not due to remarriage, but that the husband competed with another transporter named Mikkelsen, and therefore changed his surname to a patronymic based on his father Thorvald Mikkelsen. Divorces were rare, but due to remarriages, Ida Mathilde was far from unique. By linking the women using first names, date of birth and place of birth code, we can see how unusual it was to keep the surname. In Sandefjord in 1920, we found that only five out of 58 married women had the same surname as they had in 1910, and this was because they married men already holding the same surname.

6.3.3 PLACE OF BIRTH

The "Place of birth" was usually given at the municipality level – the parish had been abandoned as a census unit in the late 19th century but was still used in the church records. Spelling inconsistencies make standardization necessary, so all census records are equipped with a four-digit municipality code.⁸ Ole Kristian Kristoffersen (PID pf01036504009121) born 1848 in Sandar municipality was easy to link from the 1920 to the 1910 census using a consistent name and date of birth, but the place of birth according to the first-mentioned source created doubts, he was allegedly born in Veøy in Nordmøre, north of Bergen. Figure 7 shows that this was not a straightforward transcription error, rather, Statistics Norway's standardization of the information interpreted the farm name "Westad" as belonging on the west coast of Norway ("Veøy N Møre") and did not take into account that the same farm name exists in Sandar municipality, surrounding Sandefjord. Thus, a proximity principle needs to be applied.

A typical misinterpretation on the part of Statistics Norway concerns Kvelle, which is routinely placed in neighbouring Hedrum municipality with its sub-parish of that name, while earlier censuses placed Augusta Larsen as born in 1857 in Kville, Sweden. In contrast, there is an interesting but uncertain linkage candidate in the 1875 census for Aker bordering on Oslo, where there is a Gustav Adolf Olsen (PID pf01073803007686) born 1864 in Fredrikshald. When searching the Digital Archive, a more reliable candidate appears: Sausage maker Gustav Alfred Olsen born 1871 in "Vorter Aker" had fathered a child in Sandefjord in 1920. "Vorter" is not Norwegian, and inspection of the scanned edition of the church book shows "Vestre (Western) Aker", but the handwriting is indistinct, making it difficult to decipher this birthplace for the transcribers from India or China.⁹ The church records are often scanned from microfilm copies, which results in poorer legibility than the direct scanning of the original forms.

Figure 7 Statistics Norway violating the proximity principle by defining the farm name to be located in the western part of Norway

1. Fullt navn: Ole Christian Christoffersen

2. Mannkjønn¹. Kvinnekjønn¹.

3. Fødselsdag 15-7 i året 1848

4. Fødested: Westad Veøy Nordmøre
 { Opgi herred eller by i Norge
 eller fødeland utenfor Norge.

8 This coding of the 1920 census is done by Gunnar Thorvaldsen for the town of Sandefjord, as the first municipality in Norway in preparation for this article.

9 Birth register for Sandefjord parish 1916–1932 (0706Q). The born and baptized 1920, page number 63, line number 40.

7 CONCLUSION

In half a century, work on microdata in Norway has grown from activity in two universities transcribing and researching the full count 1801 census and a collection of census and vital records microdata from around the capital. Today, most census and ministerial records from 1801 until the mid-20th century have been scanned, transcribed, coded and made available via the websites of the National Archives and the UiT The Arctic University of Norway. Encoded and interlinked census records are also available from Minnesota Population Center as part of ipums.org. Many master and doctoral theses as well as research articles have been written on topics within social history and historical demography. Presently, research and administrative partner institutions are building the Historical Population Register with prolonged support from the Norwegian Research Council. This will contain longitudinal records of the nine million persons who lived in Norway since 1800. The register will make it possible to follow the entire population in Norway for up to seven generations where we previously could only follow samples in a few municipalities for a shorter period. Unique personal IDs with corresponding URLs to the person page provide links to many sources and introduce a superior level of historical documentation. Cross-sectional and vital records are being interlinked with automatic and manual record linkage software. Longitudinal data is available for searching as timelines and in Intermediate Data Structure format from UiT and for searching at Histreg.no, which also caters for interactive editing. Much record linkage and quality assurance work remain but we are well on the way to creating a database that can fill the void in the two centuries before the Central Population Register starts in 1964.

ACKNOWLEDGEMENTS

We thank the editors of the special issues of *Historical Life Course Studies* for their comments and good advice during our writing of this article.

REFERENCES

- Abramitzky, R., Boustan, L. P., & Eriksson, K. (2012). Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. *American Economic Review*, 102(5), 1832–1856. doi: [10.1257/aer.102.5.1832](https://doi.org/10.1257/aer.102.5.1832)
- Abramitzky, R., Boustan, L. P., & Eriksson, K. (2013). Have the poor always been less likely to migrate? Evidence from inheritance practices during the age of mass migration. *Journal of Development Economics*, 102(C), 2–14. doi: [10.1016/j.jdeveco.2012.08.004](https://doi.org/10.1016/j.jdeveco.2012.08.004)
- Alhaug, G. (2011). *10 001 navn: Norsk fornavnleksikon* [10 001 names: Norwegian encyclopedia of first names]. Oslo: Cappelen Damm.
- Alter, G. (2021). Reflections on the Intermediate Data Structure (IDS). *Historical Life Course Studies*, 10, 71–75. doi: [10.51964/hlcs9570](https://doi.org/10.51964/hlcs9570)
- Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies*, 1, 1–26. doi: [10.51964/hlcs9290](https://doi.org/10.51964/hlcs9290)
- Andresen, A. (1994). *Tromsø gjennom 10000 år, Bind 2: Handelsfolk og fiskerbønder 1794–1900* [Tromsø during 10,000 years, Volume 2: Merchants and fishing farmers 1794–1900]. Tromsø: Tromsø commune.
- Ballo, J. G. (2019). Microdata.no: New technology allows instant access to Norwegian register data. *Tidsskrift for samfunnsforskning*, 60(4), 398–408. doi: [10.18261/issn.1504-291X-2019-04-04](https://doi.org/10.18261/issn.1504-291X-2019-04-04)
- Bjørklund, I. (1985). *Fjordfolket i Kvæningen: Fra Samisk samfunn til Norsk utkant, 1550–1980* [The fjord people in Kvæningen: From Sami society to Norwegian outskirt, 1550–1980]. Tromsø: Universitetsforlaget.
- Brosveet, J., Olaussen, T. G., & Sande, T. (1979). *Kommuneendringer 1838–1978* [Municipality changes 1838–1978]. Bergen: NSD rapporter.
- Engelsen, R. (1983). Mortalitätsdebatten og sosiale skilnader i mortalitet [The mortality debate and social differences in mortality]. *Historisk Tidsskrift*, 62(2), 161–202.

- Fure, E. (1990). Personnavn og tidsånd [Personal names and the spirit of the times]. *Namn og Nemne*, 7, 35–55.
- Fure, E. (2000). Interactive record linkage. The cumulative construction of life courses. *Demographic Research*, 3, 1–20. doi: [10.4054/DemRes.2000.3.11](https://doi.org/10.4054/DemRes.2000.3.11)
- Fure, E. (2004). ... *En besynderlig regelmæssighed: Dødeligheten i Asker og Bærum på 1800-tallet med særlig vekt på spedbarnsdødeligheten* [... A curious regularity: Mortality in Asker and Bærum in the 19th century with particular emphasis on infant mortality] (Doctoral dissertation). Oslo: University of Oslo.
- Haavet, I. E. (1982). *Avvik eller uhell? Ugifte foreldre omkring 1800 — En sosial analyse*. [Deviation or accident? Unmarried parents around 1800 — A social analysis] (Master thesis). Bergen: University of Bergen.
- Hauglid, A. O. (1981). *Balsfjorden og Malangens historie* [The history of Balsfjord and Malangen]. Storsteinnes: Balsfjord kommune.
- Helland, A. (1876–1917). *Norges land og folk: Topografisk-statistisk beskrevet* [Norway's land and people: Topographically-statistically described]. Kristiania: Aschehoug.
- Holden, L., Boudko, S., & Thorvaldsen, G. (2020). Lenking og kobling i Historisk Befolkningsregister [Record linkage and family pointers in the Norwegian Historical Population Register]. *Heimen*, 57(3), 216–229. doi: [10.18261/issn1894-3195-2020-03-04](https://doi.org/10.18261/issn1894-3195-2020-03-04)
- Hovland, E. (1977). *Folket, bygda og historia* [Population, community and history]. Bergen: Universitetsforlaget.
- 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: University of Oslo.
- Kjeldstadli, K. (1987). Hva er en bydel [What is a city part?]. *St. Hallvard*, 2, 5–15.
- Kjelland, A. (2018). Mapping and analysing remigration based upon Norwegian farm- and genealogical history projects. *Journal of Migration History*, 4(2), 314–329. doi: [10.1163/23519924-00402005](https://doi.org/10.1163/23519924-00402005)
- Langholm, S. (1974). Historie på individnivå. Omkring Ullensaker-undersøkelsen — Et mikrohistorisk eksperiment [History at the individual level. About the Ullensaker investigation — A microhistorical experiment]. *Historisk Tidsskrift*, 53, 243–272.
- Langholm, S. (1976). On the scope of micro-history. *Scandinavian Journal of History*, 1(1–4), 3–24. doi: [10.1080/03468757608578894](https://doi.org/10.1080/03468757608578894)
- Lunden, K. (1975). Potetdyrkinga og den raskare folketalsvokst i Noreg frå 1815 [Potato cultivation and the faster population growth in Norway from 1815]. *Historisk Tidsskrift*, 54, 275–313.
- Naeseth, G. B., & Hedberg, B. (1993–2008). *Norwegian immigrants to the United States: A biographical directory, 1825–1850* (Vols. 1–5). Madison, WI: Vesterheim Genealogical Center and Naeseth Library, c1993-.
- Nygaard, L. (1995). *Histform — Norsk standard for registrering og utveksling av nominative folketellingsdata for årene 1865-1910* [Histform — Norwegian standard for transcription and exchange of nominative census data for the years 1865-1910]. Tromsø: Registreringsentral for historiske data, Universitetet i Tromsø.
- Olstad, F. (1995). *Sandefjords historie. Strandsitter og verdensborger* [The history of Sandefjord. Coastal resident and citizen of the world] (Vol. B. 1). Sandefjord: Sandefjord kommune.
- Quaranta, L. (2021). Reflections on the use of the Intermediate Data Structure (IDS) in historical demographic research. *Historical Life Course Studies*, 10, 76–80. doi: [10.51964/hlcs9571](https://doi.org/10.51964/hlcs9571)
- Quaranta, L., & Sommerseth, H. (2018). Introduction: Intergenerational transmissions of infant mortality using the Intermediate Data Structure (IDS). *Historical Life Course Studies*, 7, 1–10. doi: [10.51964/hlcs9288](https://doi.org/10.51964/hlcs9288)
- Rokkan, S., & Valen, H. (1966). Archives for statistical studies of within-nation differences. In S. Rokkan (Ed.), *Data archives for the social sciences* (pp. 112–127). Paris: Mouton.
- Ruggles, S., Fitch, C. A., & Roberts, E. (2018). Historical census record linkage. *Annual Review of Sociology*, 44, 19–37. doi: [10.1146/annurev-soc-073117-041447](https://doi.org/10.1146/annurev-soc-073117-041447)
- Schneider, B. R. (1974). *Travels in Computerland: Or, incompatibilities and interfaces: A full and true account of the implementation of the London Stage Information Bank*. Reading, MA: Addison-Wesley Publishing Company.
- Sogner, S., Fure, E., & Randsborg, H. B. (1984). *Fra stua full til tobarnskull: Om nedgangen i barnetall i norske familier i de siste 200 år, med særlig vekt på perioden 1890-1930* [From a housefull to parity two: On the decline in child numbers during the last 200 years, emphasizing 1890–1930]. Oslo: Universitetsforlaget.

- Sogner, S., Randsborg, H. B., Fure, E, & Walloe, L. (1986). Le déclin de la fécondité en Norvège (1890–1930) [The decline in fertility in Norway (1890–1930)]. *Annales de Démographie Historique*, 1986, 361–375. doi: [10.3406/adh.1987.1669](https://doi.org/10.3406/adh.1987.1669)
- Sommerseth, H. (2011). *Northern co-residence across generations. In northernmost Norway during the last part of the nineteenth century* (Ph.D. thesis). Tromsø: UiT The Arctic University of Norway. Retrieved from <https://munin.uit.no/bitstream/handle/10037/3372/thesis.pdf?sequence=5&isAllowed=y>
- Sommerseth, H. (2018). The intergenerational transfer of infant mortality in northern Norway during the 19th and early 20th centuries. *Historical Life Course Studies*, 7, 69–87. doi: [10.51964/hlcs9284](https://doi.org/10.51964/hlcs9284)
- Sommerseth, H., & Thorvaldsen, G. (2022). The impact of microdata in Norwegian historiography 1970 to 2020. *Historical Life Course Studies*, 12, 18–41. doi: [10.51964/hlcs11675](https://doi.org/10.51964/hlcs11675)
- Statistics Norway. (1880). *Statistik angaaende det Norske jordbrug. Fornemmelig i femaarsperioden 1871–1875 og i året 1875* [Statistics on the Norwegian agriculture. Especially 1871–1875]. Kristiania: Trykt i Ringvolds bogtrykkeri.
- Statistics Norway. (1980). *Folketeljinga 1801. Ny bearbeiding* [Population census 1801. New processing]. Oslo: Statistisk sentralbyrå.
- Statistics Norway. (1995). *Historisk statistikk 1994* [Historical statistics 1994]. Oslo: Statistisk sentralbyrå.
- Thorvaldsen, G. (1992). The preservation of computer readable records in the Nordic countries. *History and Computing*, 4(3), 211–227.
- Thorvaldsen, G. (1994). The encoding of highly structured historical sources. *Computers and the Humanities*, 28, 301–305. doi: [10.1007/BF01830278](https://doi.org/10.1007/BF01830278)
- Thorvaldsen, G. (1995a). Longitudinal sources and longitudinal methods — Studying migration at the Stockholm Historical Database. In A. Brändström & L.-G. Tedebrand (Eds.), *Swedish urban demography during industrialization* (pp. 219–250). Umeå: Umeå University.
- Thorvaldsen, G. (1995b). *Migrasjon i Troms i annen halvdel av 1800-tallet. En kvantitativ analyse av folketellingene 1865, 1875 og 1900* [Migration in Troms province in the second half of the 19th century. A quantitative analysis of censuses 1865, 1875 and 1900] (Doctoral dissertation). Universitetet i Tromsø, Tromsø. Retrieved from <https://hdl.handle.net/10037/1390>
- Thorvaldsen, G. (1996). *Håndbok i registrering og bruk av historiske persondata* [Handbook on transcription and use of historical person data]. Oslo: Tano Aschehoug.
- Thorvaldsen, G. (1997). On boundaries and areas in local historical research. In J. E. Myhre & K. Kjeldstadli (Eds.), *Festskrift til Sivert Langholm*. Oslo: The Norwegian Historical Association. Retrieved from <https://hdl.handle.net/10037/29041> (web version in English)
- Thorvaldsen, G. (2006). Away on census day. Enumerating the temporarily present or absent. *Historical Methods*, 39(2), 82–96. doi: [10.3200/HMTS.39.2.82-96](https://doi.org/10.3200/HMTS.39.2.82-96)
- Thorvaldsen, G. (2008). Hushjelper og jordbrukstjenere — Når kom nedgangen i tjenerallene? [When did the number of domestic servants decline in Norway?]. *Historisk Tidsskrift*, 87(3), 451–464. Retrieved from <https://hdl.handle.net/10037/2014>
- Thorvaldsen, G. (2011a). Household structure in the multi-ethnic Barents region: A local case study. In D. G. Anderson (Ed.), *The 1926/27 Soviet polar census expeditions* (pp. 117–132). Oxford, New York: Berghahn.
- Thorvaldsen, G. (2011b). Using NAPP census data to construct the Historical Population Register for Norway. *Historical Methods*, 44(1), 37–47. doi: [10.1080/01615440.2010.517470](https://doi.org/10.1080/01615440.2010.517470)
- Thorvaldsen, G., & Erikstad, M. (2007). Utvandring til Sverige på 1800-tallet [Emigration to Sweden in the 19th century]. *Heimen*, 44(2), 117–130.
- Thorvaldsen, G., Pujadas-Mora, J., Andersen, T., Eikvil, L., Lladós, J., Fornés, A., & Cabré, A. (2015). A tale of two transcriptions. Machine-assisted transcription of historical sources. *Historical Life Course Studies*, 2, 1–19. doi: [10.51964/hlcs9355](https://doi.org/10.51964/hlcs9355)
- Thorvaldsen, G., Sommerseth, H. L., & Holden, L. (2020). Anvendelser av Norges historiske befolkningsregister [Interfaces to Norway's historical population register]. *Heimen*, 57(3), 230–243. doi: [10.18261/issn.1894-3195-2020-03-05](https://doi.org/10.18261/issn.1894-3195-2020-03-05)
- Vick, R., & Huynh, L. (2011). The effects of standardizing names for record linkage: Evidence from the United States and Norway. *Historical Methods*, 44(1), 15–24. doi: [10.1080/01615440.2010.514849](https://doi.org/10.1080/01615440.2010.514849)
- Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association* (pp. 1–8). Retrieved from <https://eric.ed.gov/?id=ED325505>

HISTORICAL LIFE COURSE STUDIES
VOLUME 12 (2022), published 20-06-2022

The Groningen Integral History Cohort Database

Development, Design and Output

Richard Paping

University of Groningen

Dinos Sevdalakis

University of Groningen

ABSTRACT

The Groningen Integral History project launched in 1987 aimed to sketch the lives of people from various social classes in the Dutch province of Groningen in the 19th and early 20th century. One part was the creation of the Groningen Integral History Cohort Database (GIHCD), reconstructing complete individual life courses of 5,280 persons (RPs) born between 1811 and 1872. The quality of the database has become very high by now, despite the lengthy and difficult process of shaping it over 35 years. More than 98% of the RPs (and for some parts of the database even more than 99%) could be followed until their death or until a migration abroad. Even for the life courses of those moving abroad information is available for most RPs. In this article, we primarily focus on the rural part of the database (n = 4,320), the quality of which is the highest and has had the most significant tangible research impact. Building on information from the Dutch civil registration system (from 1811) and the population registers (from 1850), the database includes multiple individual-level variables. In the technical part of the article, we provide an extensive overview of the available variables and summarize the transformation of the rural part of the database into an Intermediate Data Structure (IDS). Since the early 1990s, historians from the University of Groningen have used GIHCD in quite some publications. At the end of this article, we provide a summary of the main outcomes of these publications.

Keywords: Demography, Historical databases, Life courses, Groningen, Dutch Civil registration, IDS

DOI article: <https://doi.org/10.51964/hlcs12033>

© 2022, Paping, Sevdalakis

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

The Groningen Integral History Cohort Database (GIHCD) consists of a regional sample of more than 5,000 life courses of persons born between 1811 and 1872 in the northern Dutch province of Groningen. This province consists of a large city (Groningen) and its surrounding countryside with numerous smaller and larger settlements organised in about 60 municipalities in the 19th century. Most of the individuals involved were followed during their life throughout the Netherlands, and largely also abroad until well into the first half of the 20th century. The database's main strength is its excellent quality, as a very high percentage of the included persons could indeed be followed throughout their entire life course.

This article first presents an overview of the history of the Integral History Project out of which the GIHCD developed from 1987 onwards. Secondly, we describe the rather unsystematic way in which the database was constructed over the years, partly in response to changing research questions. In the next section, we discuss what the database looks like at the moment, presenting its content and structure. In the last part, we will present an overview of the publications that made use of the data stored in the database. In general, the article also provides an illustration of how technological developments over more than thirty years, in addition to a continuous lack of time and money, can shape a database like GIHCD into its present form.¹

2 THE ROOTS OF THE DATABASE

2.1 THE START OF THE INTEGRAL HISTORY PROJECT

In 1987, researchers from both the University of Groningen (Pim Kooij and Marten Buist) and the University of Utrecht (Gerard Trienekens and Theo van Tijn), with support from the organisation for historical research Stichting voor Historisch Onderzoek (SHO), one of the predecessors of the government-funded Dutch organisation for scientific research NWO, launched the Integral History Project (Kooij, 1993a, p. 2). The primary goal of the project that led to the creation of the GIHCD, was to write an integral history of different regions in the north and the south of the Netherlands. Through this approach the initiators sought to solve the ongoing fragmentation of historical research by operationalising the concept of "quality of life" (Kooij & Sleebe, 1991; Trienekens, 1987, 1993). In short, the theoretical idea was to compare the aims of ordinary people regarding various aspects of their lives with their perception of what was happening in "reality", thus enabling historians to research the quality of life of people in the past. According to this reasoning, major differences between ambitions and perceptions of people will have led to social tensions that would clearly show up in the sources. As general developments — for instance technological and economic progress, modernisation in other respects and democratisation — influenced these aims and perceptions, this would also result in opportunities to connect data on micro and macro level.

Although at first sight events were not the core of the Integral History Project, the main way to trace indications of changes in ambitions and perceptions was by doing extensive and detailed archival research on specific individuals and communities representing different societies. Four research pillars were defined, that had to result in a similar number of large databases: 1. Cohort analysis: research on the life course of the inhabitants of selected municipalities; 2. Structural analysis: research based on snapshots of the household structure of the inhabitants in these selected municipalities taken every 20 years; 3. Financial analysis: research on the developments in the provincial and municipal expenditures using the annual accounts (Duijvendak & Blijham, 1994) to study the political priorities of the local and provincial elites; 4. Opinion analysis: research based on the local newspapers to find indications of tensions in society. For this article, we concentrate on pillar 1 and to some extent on pillar 2, because they at least have had some long-lasting research output.

Originally, two Dutch cities and their surrounding regions were selected to be studied in the period 1770–1914. On the one hand, the city of Groningen and the clay soil region situated largely north and east of it, and on the other hand, eastern Northern Brabant with the capital of Den Bosch (van Oudheusden & Trienekens, 1993). So, next to the capital, nine more or less rural municipalities were chosen to be investigated in detail for both regions.

Regarding the Northern Brabant part of the project, it must be noted that it was attempted to reconstruct the life courses (pillar 1) engaging many volunteers. However, this data collection — partly due to unsolved

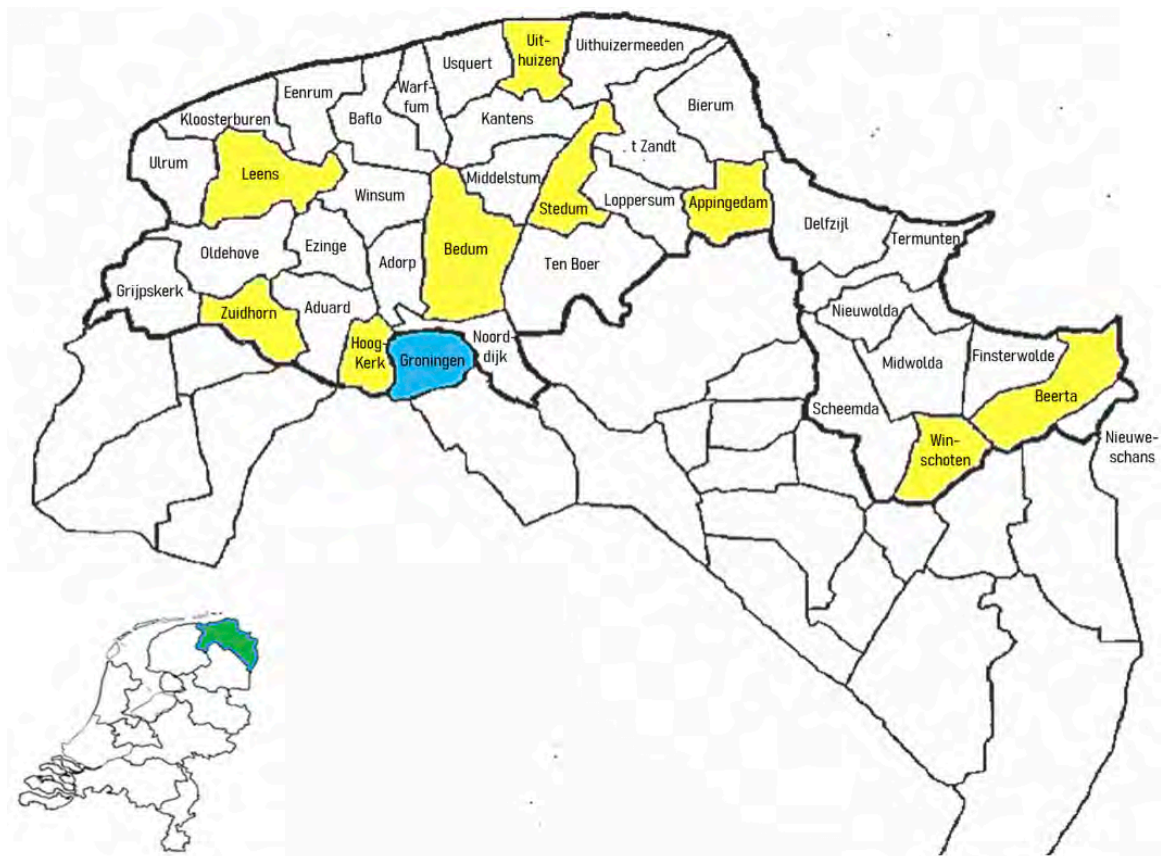
1 The authors want to thank Geurt Collenteur, Maarten Duijvendak, Frouke Hansum, Kees Mandemakers and H el ene V ezina for their many useful comments.

technical problems — did not result into a database that could be analysed easily. Still, some results from the Brabant cohort analysis were published in Schrover (1998). Largely due to the different sample years chosen for each municipality, the collection of the household structures (pillar 2) turned out to be problematic as well. Preliminary results were presented by Trienekens (2004), while a database called "Het Huishouden in Oostelijk Noord-Brabant 1800–1920" is available.² However, the overall output of the Northern Brabant part of the project is limited. Consequently, this article will concentrate on the Groningen part of the project.

The Groningen clay region was largely defined as those municipalities whose surface area was covered by clay for more than 50% (Paping, 1995, pp. 18–21). The selected region covers more than half the province including 36 out of the 57 19th-century municipalities (see figure 1). The nine municipalities were chosen in such a way to ensure an even geographical spread over the region. They were largely rural, though also deliberately comprised of the small old town of Appingedam with its adjoining rural villages and the somewhat larger semi-urbanised settlement of Winschoten. However, for the sake of convenience, we will call these nine municipalities rural in opposition to the large city of Groningen.

The Dutch scientific fund SHO, in addition to financing PhD's, provided the Integral History Project in 1987 with some funding to develop the four databases. In the Groningen part of the project, this money was used partly to finance a fulltime research assistant, who constructed a digital database of government accounts (pillar 3). However, the amount of funds supplied by SHO were by no means sufficient to finance the construction of the four large historical databases, just mentioned. The University of Groningen also funded additional PhD's and student-assistants to participate in the project.

Figure 1 *The municipalities of the clay region in the Groningen province*



Source: Paping (1995).

Explanation: The so-called 'clay area' is the area surrounded by the thick line and the named municipalities. The municipalities in yellow and blue are the ones that are included in the GHICD database.

2 See <https://www.narcis.nl/research/RecordID/OND1296249/Language/nl>

2.2 THE STRUCTURAL ANALYSIS DATABASE

The aim of the structural analysis was to take snapshots of the household situation every 20 years from 1815 onwards as benchmarks to study changes over time. The transcription and summarising of the population registers of the nine selected rural municipalities was done by several persons working on PhD-projects, some supported by student-assistants. Originally, a similar endeavour was planned for the city of Groningen for the period before 1870, as the years 1870–1910 had already been covered in the thesis of Kooij (1987). However, the urban part of the project was never completed.

The structural analysis was confronted with three major problems. Firstly, there was a lack of sources. The intended timespan of the project was 1770–1920; however, population registers or micro census-data were missing for the older period. Therefore, it was decided to restrict the data collection to the census years 1815, 1829/1830, 1850, 1870, 1890 and 1910, reconstructing the household situation around January 1st of those years. Unfortunately, for most of the selected municipalities the (census) registers were missing for 1815 and 1830. From 1850 onwards, nationwide introduced dynamic population registers were used, except two municipalities for which these registers were missing for the first decade. Secondly, it often proved difficult to reconstruct the exact household situation in 1870, 1890 and 1910 as in the countryside many of these dynamic population registers cover periods of twenty years or more, for instance from 1860 to 1900, while the original ten-yearly census forms had been destroyed. This made it very difficult to ascertain which persons (including live-in servants) lived within the households at the measure points from 1870 onwards. Thirdly, the reliance on only a few PhD-candidates — due to the lack of funding — meant that the years 1870, 1890 and 1910 were only done to a limited extent. Consequently, only for 1830 and 1850 a digital database on household structures is available for a substantial number of municipalities (see table 1).

As laptops and the like were still rare at the end of the 1980s, all households were registered on a paper form. Standardly, all data of the head of the household and eventual partners were transcribed including the names. For the other members of the household some data were transcribed as well, though in much less detail than available in the registers in an attempt to reduce the time needed to collect the data. After filling out all the forms, the information was translated into numbers using an extensive codebook (with numbers indicating, for instance, household types, occupations, and so on). These numbers were digitalised on tape and analysed by way of SPSS using a mainframe computer. Around 1995, a LOTUS 1-2-3 spreadsheet database was created (later on converted into EXCEL) with a household on every row. The numerical codes were reconverted into the corresponding meaning in text. Next, some of the not yet digitalised information from the still preserved original forms was added to the database. An important feature of this dataset is that for more than half of the municipalities in 1830 and 1850 information on the amount paid in the local taxes could be added (the so-called 'hoofdelijke omslag', a tax based partly on income and partly on wealth: Kooij, 1993b, pp. 145–155; Paping, 2010; Voerman, 2001, pp. 267–272).

Table 1 *Overview of the available databases regarding household structure constructed for the nine rural municipalities of the Groningen Integral History Project*

	1830	1850	1870	1890	1910
Appingedam	Missing	C/D	C	C	C
Bedum	C/D/T	C (partly)	C/D/T	C/D	C/D
Beerta					
Hoogkerk	Missing	C/D/T	C/D/T	C/D/T	C/D/T
Leens	C/D/T	Missing			
Stedum	C/D	C/D			
Uithuizen	Missing	C/D/T	C	C	C
Winschoten	C/D	C/D/T	C/T	C/T	C/T
Zuidhorn	Missing	C/D/T		C/D	C/D

Explanation: C: Reconstructed snapshots; D: Digitalised and standardised databases; T: Tax information available. Empty cells: sources are available, though no databases have been reconstructed.

The structural analysis databases were used for several publications: Voerman (2001) made an in-depth analysis for Winschoten in combination with extra data on the mixed rural-urban municipalities of Hoogezand and Veendam in the eastern Groningen peat districts neighbouring the Groningen clay soil region. Paping and Collenteur (1998) used it for the long-term development of the occupational structure of household heads until 1910. Paping (2010) researched the relation between occupation, social position and tax performance within households in the period 1830–1850; and lastly Paping (2018) analysed the household structure using the same data.

2.3 THE CONSTRUCTION OF THE GIHCD

This article focuses in particular on the most prominent database constructed within the Groningen Integral History Project: the cohort database with individual life courses from ten municipalities (including the city of Groningen, see figure 1), selecting 120 births from cohorts starting with a 20-yearly interval (1811, 1830, 1850, 1870 and 1910). Although the original aim included the birth cohorts of 1770 and 1790, it soon became clear that this was not feasible. For the three villages later constituting the municipality of Hoogkerk, the first 120 baptisms for the cohort of 1770 and 1790 were collected by Vincent Sleebe (1993). However, it proved extremely difficult to find any other information about these persons. In Groningen, the large majority of the families did not have a surname before 1811, making it difficult to trace persons outside the village without extensive genealogical research. Initially, it was hoped that existing registrations of protestant church members (*lidmaten*) would indicate migrations, but these registers proved scarce or incomplete and covered only a minority of the population. This was even more detrimental, as the naïve initial research assumption that migration was very rare in the countryside proved to be completely beyond reality. How to deal with migrating cohort members became a persistent issue in the project as we will show later on.

The project originally planned a series of books on all 10 municipalities involved starting with Hoogkerk, a tiny municipality right next to the city of Groningen. This characterised the optimistic atmosphere in the first years of the project: ultimately only the book on Hoogkerk (Kooij, 1993) was published. The contribution to this book that makes extensive use of the GIHCD clearly shows the huge problems with which the construction of the cohort analysis was confronted (Clement, 1993). Birth cohorts were constructed for 1811, 1830, 1850 and 1870. As the population of Hoogkerk was very limited, it sometimes took even six years to achieve a cohort of 120 consecutive births. The cohort of 1811 started in August, with the beginning of the official Dutch civil registration. The birth certificates offered much more information (occupations and ages of people involved) than the baptism registers which had to be used before August 1811. At first, children were only traced within the municipality of birth and in the city of Groningen. As most of the Hoogkerk children disappeared without a trace, this search was expanded to neighbouring rural municipalities like Aduard, where indeed a few of the Hoogkerk cohort members were found.

In the 1980s, it was still difficult to track down a specific person in the civil registration. Only 10-yearly indices of births, deaths and marriages existed for each municipality, supplying merely the registration date of the certificates. As it was not uncommon to have several people with the same name, all hits had to be checked one by one. Fortunately, this could be done at the provincial archive, where an increasing amount of films could be found containing the civil registration of the provincial municipalities before 1900. A notable exception were the records of the city of Groningen, which were kept in the city archive. In 2002 the two archives fused (Groninger Archieven) and the work could be concentrated in one building.

Originally, the core of the database was formed by standardised paper forms for each Research Person (RP). These forms had room for information about: 1. The parents (occupation, birth date, birth place, ability to sign birth or marriage certificate); 2. Marriage (date and location) and characteristics of the partner (occupation, birth date, birth place, ability to sign marriage certificate); 3. Children (birth date, occupation and moment they left the parental household, including stillbirths); 4. The religion, the household situation and exact place where the RP lived according to the census (1815, 1830 and 1840) or the population register (from 1850 onwards); 5. Migration dates and destinations. Extra forms could be added when necessary. Consequently, the original database contains information on three generations, not only on the RP, but also on his or her parents, on marriage partners and on the place and date of birth of children.

The advantage of the paper system was its flexibility, as it was easy to include interesting additional information found in archives. The disadvantages were, firstly, that it was often unclear what the source of the transcribed information was and, secondly, that forms were being filled in very inconsistently. The form gave a lot of decision power to the individual researcher about which information to be included. Consequently, a lot of data was not transcribed, making it often necessary to go back to the original sources and replenish the data

on the forms. Most of the forms were filled in by student assistants and research assistants, though also sometimes by history students as part of a paper they wrote during a research course under the supervision of project leader Pim Kooij. Kooij himself did most of the cohorts of the city of Groningen. The handwriting, and also the colour of the pen, pencil or marker provide indications on who originally transcribed the data. Over the course of time, many different researchers worked on one individual RP.

Before 1995, the tracing process in the civil registration proved extremely difficult and time consuming. Especially the high number of persons in the province of Groningen sharing the same name made it time-consuming to establish if a certificate belonged to the RP in question or not, without reading the certificate itself. In 1993/1994 all funds of the Integral History Project had been used. All life courses from the cohorts from 1811 until 1870 had been constructed, except migration routes outside the birthplace. A few student-assistants — financed by the Groningen Faculty of Economics — had tried to improve this; however, it was too much work. In this phase, RPs were followed in the municipalities in the clay parts of the province, and not only in the city of Groningen. During this period, project leader Kooij created a REFLEX-database with a structured summary of the assembled data of every individual based on the paper forms. In Figure 2, we sketch the technical transformations the database has undergone since 1987.

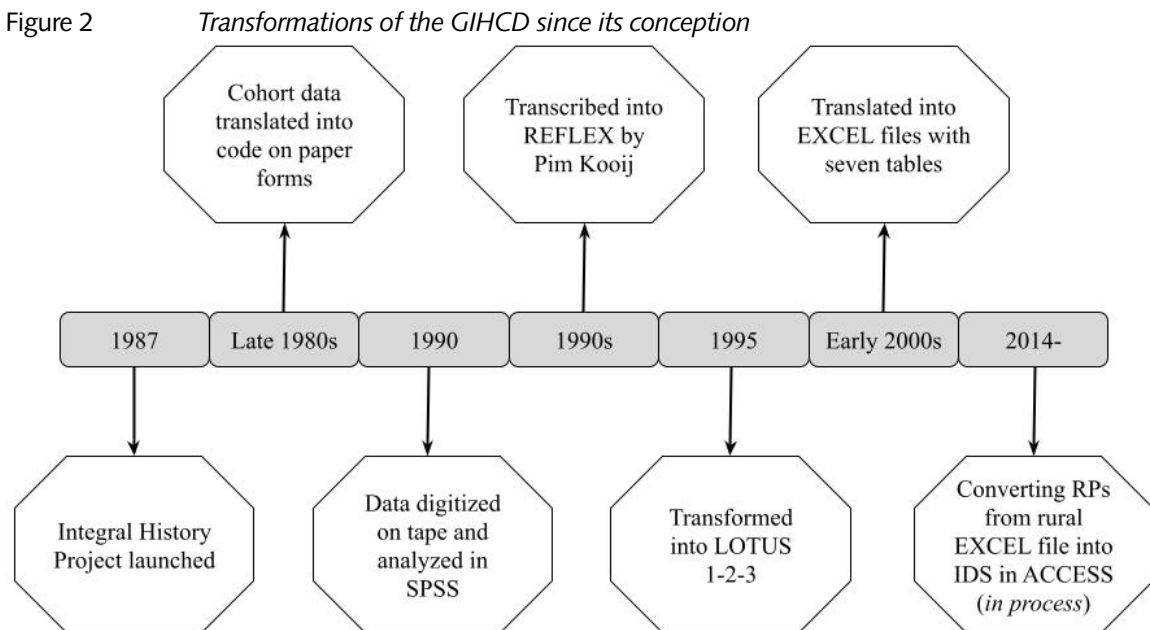
The Integral History Project received some fresh money from the Dutch research foundation (NWO) to stimulate research cooperation with historians from the former USSR. Since the end of the Communist regime, there was a rising interest in Western research methods among Russian historians, together with a change of focus from political history (mainly the history of communism) to the history of common people. This Dutch-Russian cooperation resulted in several workshops in both Russia and the Netherlands and in two volumes: *Where the Twain meet* (Kooij, 1998) and *Where the Twain Meet Again* (Kooij & Paping, 2004). Several Russian contributions applied the Groningen method of cohort analysis (Akolzina et al., 2004; Dyatschkov et al., 1998; Golubeva, 1998; Shustrova & Sinitsyna, 2004; Sinitsyna, 1998). However, source problems (serious gaps in birth, marriage and death records and difficulties with record linkage) were even larger for 19th century Russia, resulting in databases containing only a limited number of individuals with substantial information.

In the meantime, NWO funded in the years 1995–1996 a Dutch-Flemish research project concentrating on family strategies. Paping's part of the project used the Groningen Integral History Cohort Database (GIHCD) to study strategies regarding migration in the countryside. He had to conclude that — despite all the efforts — the cohort analysis database was still not very consistent and had significant gaps. The quality of the 1811 cohort was especially poor, partly due to the difficulties to detect the migration history of the cohort members, as the Dutch population registers keeping systematic track of migration only started in 1850.

As a preparation for the family strategy project, the original REFLEX-database was converted into a LOTUS 1-2-3 spreadsheet, making the data analysis and the input of extra data easier. However, one major weakness of the used program was that it had difficulty accepting dates before 1880. To solve this problem, 100 years were added to all the dates. In the end, dates in the LOTUS database were changed into relative parts of the year — for instance, the 30th April 1840 became 1840.33 — making it easy to calculate time spans. This implied that the exact dates got lost, although these are still available on the original forms. Next, all the data of all RPs from the nine rural municipalities and the cohorts 1830, 1850 and 1870 were checked by Paping, focussing on RPs with an incomplete life course. It turned out that previous investigators had missed a lot of information, which sometimes could even be found at the place of birth. For the still numerous missing persons, it was checked if they had gone to the city of Groningen or if they appeared on American migration lists. Also, many lost cohort members were found by using newly available indices and films of civil certificates and by systematically scrutinising the whole collection of published genealogies in the provincial Groningen Archive. In general, cohort members were followed through the whole province of Groningen at least until 1900, but preferably until after 1920.

Consequently, by 1999 the quality of the 1830, 1850 and 1870 rural cohorts had been greatly improved. Of the 1830 and 1850 cohort members only 8% and 10% were lost somewhere during their life in the province of Groningen. The situation was worse for the 1870 cohort where about a quarter had not (yet) been traced during their life span; however, only 4% disappeared before 1900 (Paping, 1999, p. 79). These first overviews of the cohorts showed a massive and rising emigration of the cohort members out of the province of Groningen, respectively 11% of the 1830 cohort members, 18% for 1850 and even 26% for 1870. These observed migration shares are even more impressive taking into account the high child and juvenile mortality — as RPs dying early obviously did not have much time to move out of the province — and the percentages of lost persons just mentioned.

Figure 2



In 2002, in line with these interesting migration patterns, the Faculty of Arts of the University of Groningen financed a research project following the migration all over the Netherlands. The research period was extended until 1940 and in first instance extra information on migrations within the province of Groningen was gathered. Subsequently, RPs were followed outside the province, by visiting the municipal archives of main destinations like Amsterdam and writing letters to the smaller archives. This research effort again greatly improved coverage and reliability of the GIHCD, which by then had been converted from LOTUS into EXCEL. However, the extremely time-consuming gathering of data on all migrated individuals, was one of the major causes for this project not to have resulted in an end publication.

Since the millennium change, digitisation of the civil registration has rapidly proceeded. Thanks to the efforts of the Groningen archives, indices with the names, ages and occupations mentioned in all birth, death and marriage records became increasingly available via the website [AlleGroningers](#) (see also [Mandemakers, Bloothoof, & Laan, forthcoming](#)). Numerous volunteers made the transcripts and improved the Groningen Archives database, adding extra years when allowed by Dutch privacy regulations (public after 100 years for births, 75 years for marriages and 50 years for deaths). A more recent development was the addition of the scans of the original sources. Since the early 2000s, a sophisticated online search engine made it possible to trace individual persons from behind the computer! It became much easier to establish if a certificate really was about the person one was looking for, and this without the time-consuming travel to any archive. One can also search for a combination of names which makes it easy to find specific couples. Also, for other parts of the Netherlands search engines became available to tackle huge databases — especially [WieWasWie](#) covering the whole of the Netherlands, though also search engines from separate archives — making it increasingly easier to track down lost cohort members.

In conclusion, digital tools have made it much easier from about 2005 onwards to develop the GIHCD. These digital improvements have been especially useful for improving the quality of the 1811 rural cohorts, which had been largely neglected since 1995, as well as the 1811, 1830, 1850 and 1870 cohorts of the city of Groningen. However, due to lacking research funds the Integral History Project was still unable to do a general update of all cohorts.

Several history students wrote their master or bachelor thesis using the GIHCD: Leendert Klokkenburg (2009) discussed differences between orthodox and non-orthodox Calvinist rural RPs, Bart Hoogenboom (2013) investigated migration from the Groningen countryside to the city of Groningen and Piet-Jan Koning (2019) researched the huge migration from the Groningen countryside to the United States of America (USA). Koning showed that it is possible to trace at least three quarter of the migrating rural RPs born in 1850 and 1870 in America using among others the digitalised US census and passenger list data available online. The relatively small sample size allowed Koning to search for these lost RPs by hand and ultimately led to the retracing of 152 of 219 RPs studied that moved to the US. This seems better than software-based approaches used in recent initiatives to match and validate almost 500 persons of the HSN who migrated to the USA using American censuses (see [Paiva, Anguita, & Mandemakers, 2020](#)). So, retracing lost RPs in the USA manually might be a fairly efficient project when conducted on a small scale.

Another stimulus for the GIHCD was its participation in the European Historical Population Sample Network (EHPS-Net) in the period 2011–2016. Jacek Pawlowski was hired in 2014–2015 to put the data of the 1830, 1850 and 1870 RPs in the Intermediate Data Structure (IDS; [Alter & Mandemakers, 2014](#)). In this period a general update of the database took place by digitally tracing part of the remaining lost cohort members using newly available search engines.

3 CONTENT AND STRUCTURE OF THE GIHCD

3.1 OVERVIEW

As has been explained in the previous section, a major part of the Integral History Project focused on sampling microlevel data for the province of Groningen. The GIHCD sampled data on RPs born between 1811 and 1870 based on birth certificates. However, the history, the scope, the data collection and the sampling strategy was much different from that performed by the Historical Sample of the Netherlands (HSN, see [Mandemakers, 2000](#)).

The life course data is mainly, but not exclusively, drawn from civil registration documents (birth, death and marriage records) that describe individuals' life events. For the period after 1850 — when the Dutch state required municipalities to systematically keep track of their population — population registers have also been used. Ultimately, of a total of 5,280 RPs involved, information on 3,240 of them, derived from the nine rural municipalities (see figure 1), has been converted into an IDS database ([Alter & Mandemakers, 2014](#)). The dynamic nature of the Dutch population registers allowed documenting a number of characteristics of the RPs — like place of residence and occupation — throughout their lives.

As the database does not contain identifiable personal information relating to RPs younger than 100 years and the data of other persons such as children of the RPs are only anonymously processed, there are no impediments regarding privacy. At the moment, researchers can only get the database on request, though it is the aim to make it downloadable in the near future starting with the parts converted into the IDS format. This would allow researchers to use the sample for various studies on demographic and socioeconomic history, as the database includes precise and complete information on the RPs' occupations and migration history as well as some information on their family members. In section 3.2, we will describe in detail the sample design and content of the database and in section 3.3 we focus on the IDS component, which is also the most complete part. When relevant we also reflect on those parts of the database that are still compiled in an EXCEL file. An earlier description of the GIHCD can be found online.³ This description provides a detailed introduction to the data included in the database.

Besides the GIHCD, a related and partly supplementary database is also available. The Database Roman-Catholics Groningen consists of a family reconstitution of all Roman-Catholics — about 5% of the population — roughly living in the 'GIHCD-part' of the Groningen countryside in the 18th century ([Paping, 2009](#); [Paping & Collenteur, 2004](#); [Paping & Schansker, 2013](#)). This database — the first version of which was constructed for the master thesis of the first author of this article — offers systematic information on the life courses of about 5,000 persons born between 1721 and 1810.

3.2 SAMPLE DESIGN, SOURCES AND CONTENT

The GIHCD focuses on the large city of Groningen and on nine of the 56 rural municipalities that made up the province of Groningen in the 19th century. The nine municipalities are Zuidhorn, Hoogkerk, Leens, Uithuizen, Bedum, Appingedam, Stedum, Beerta and Winschoten. Together they comprise nine of 36 rural municipalities in the clay parts of the province. The selection method of the 10 research municipalities has already been discussed in section 2.1, where a map is also presented (figure 1). Table 2 shows some characteristics of these municipalities.

Using the 1862 agricultural statistics ([Bijdragen, 1870](#), see Table 2), the share of members of farm labourers' families in the total population of the clay parts of Groningen can be estimated as 41%, while its unweighted average share in the selected nine rural municipalities is only 38%. This figure is a little bit lower than the total average for the clay region, because we have included in our rural sample two relatively urbanised

3 See "Integral History Project Groningen", *EHPS-Net*: <https://ehps-net.eu/databases/integral-history-project-groningen>, assessed on 20 April, 2022.

municipalities, Appingedam and Winschoten. For members of farmers' families, the shares in total population are 19% and 20% respectively. The remaining 40% of the population consisted mainly of the families of artisans, millers, shopkeepers, innkeepers, merchants, shippers, reverends, schoolmasters and others active in industry and services. This high non-agricultural share denotes the large extent of specialisation of tasks within the regional economy, which was combined with a very market-oriented agriculture. Before the end of the 19th century, large-scale factories were rare in the clay parts of the province of Groningen (Paping, 1999).

The initial sampling of RPs was based on the birth certificates, being available from August 1811. The sample was drawn with intervals of about 20 years, which led to four cohorts of individuals born in 1811, 1830, 1850 and 1870, though some cohorts include some extra years, due to the small size of some of the municipalities involved. For every municipality, the first 120 registered births in each cohort were used. For the large city of Groningen 240 certificates were selected, using the first 20 of every month.

The selected RPs were subsequently traced along specific moments during their life course, such as their wedding, death and birth of their children. The sources which were systematically checked were the birth, marriage and death certificates and the population registers starting in 1850. Other sources have been used to corroborate existing information or for tracking a person who could not be found in the mentioned ones. Examples of these sources include the census lists of 1815, 1829/1830 and 1839/1840, genealogies, military draft records, migration lists and tax records.

Unfortunately, due to the lack of systematic household information on the forms (partly also due to the limited availability of this information in the population registers before 1862), the database does not contain data on the precise household composition over time. It only provides information on the parental or marital relationships between individuals. This allows researchers to analyse families, but not households. However, this might be less of a problem, as the dominant household form in the province of Groningen was the nuclear family, and estimates are available on the date when the RP left the parental home (Paping, 2004b, pp. 278–279; Paping, 2018).

For the cohorts that were drawn from the city of Groningen more linkages are missing (see Appendix A1–4), although recent efforts have improved them to a considerable extent. Tables 3 and 4 provide an overview of how far RPs could be followed during their life course. As the information regarding the urban cohorts has been digitalised only partly, these cohorts have been excluded from tables 3 and 4. The lower result for the 1811 cohort is partly caused by the relatively high number of inaccuracies in the 1811 birth registration, as these were the first official birth records in the new Dutch civil registration.

Table 2 *Description of the municipalities selected by the Integral History Project Groningen*

Municipality	Description	Population 31 Dec. 1829	Population 31 Dec. 1899	% farmers 1862	% farm labourers 1862
Groningen	Large urban centre	30,260	66,537	0%	0%
Appingedam	Small city & 5 hamlets	2,855	4,348	14%	15%
Bedum	Large village & 5 hamlets	2,720	4,894	29%	40%
Beerta	2 larger villages	2,777	4,123	15%	42%
Hoogkerk	3 villages	815	2,082	41%	30%
Leens	6 villages	2,736	3,902	16%	44%
Stedum	3 villages	1,354	2,257	20%	54%
Uithuizen	Large village	1,921	3,730	13%	49%
Winschoten	Urbanised village	3,229	9,668	9%	13%
Zuidhorn	2 larger villages	1,720	2,840	16%	44%

Explanation: The percentage of the members of farm labourers' families (including farm servants) and of the farmers' families in 1862 (based on data from Bijdragen, 1870) is calculated as a share of the total population of a municipality according to the census of 31 December 1859 (www.volkstellingen.nl).

Table 3 *Life course characteristics of the 1811–1870 'rural' cohorts (n = 1,080 per cohort)*

	1811	1830	1850	1870	Total
Lost	2.1%	1.8%	1.2%	0.6%	1.4%
Emigrated permanently	4.2%	8.0%	11.5%	16.7%	10.1%
Died before 20 in the Netherlands	28.0%	25.6%	29.0%	33.5%	29.0%
Died after 20 in the Netherlands	65.7%	64.6%	58.3%	49.2%	59.5%
Marriage year known	56.1%	58.4%	55.8%	54.2%	56.1%

Notes: The 1811 cohort has not been converted into IDS yet. For a data overview per municipality, see Appendix A.

Table 4 *State of the life courses in the rural part of the GIHCD, per age group for the total of the cohorts 1830, 1850 and 1870 (n = 3,240).*

	Migrated permanently	Died within the Netherlands	Lost in research process	Remaining in the database
Aged 0–20	4.1%	29.4%	0.6%	Age 20: 66.0%
Aged 20–40	5.7%	11.6%	0.5%	Age 40: 48.2%
Aged 40–60	2.0%	10.4%	0.1%	Age 60: 35.7%
Aged 60+	0.3%	35.4%	0.0%	–

Table 3 shows that only a few RPs were lost without a trace. Even the cohort of 1811 seems relatively good. So, for all cohorts most of the life course of the RPs is known. The reported huge death rate of juveniles in 1870 can be contributed to a smallpox epidemic, which substantially increased infant mortality, making the choice for this specific sample year rather unfortunate. Both the increasing child mortality and the increase in emigration — not all the marriages abroad have been tracked yet — resulted in a slight decline of the number of known marriages in the database between 1830 and 1870.

As explained in section 2.3, from 1995 onwards the research effort was primarily focused on the rural 1830, 1850 and 1870 cohorts. This effort concentrated on the manual linking of the RPs with their relatives (parents, spouse and children). By doing intensive research in online databases and other sources of information for missing information on parents, marriage, children death and migration, nearly all RPs have been successfully linked with their relatives. Table 3 shows that for the 1811–1870 cohorts, we have no death certificate or emigration data for on average 1.4% of the 4,320 RPs, ranging from 2.1% for the 1811 cohort to 0.6% for the 1870 one. For the 1830–1870 cohorts this is even only 1.2% (table 4). Including also parents, husbands and children the last cohorts contain information on 19,045 individuals.

Recently collected information on the fate of most of the RPs who emigrated to the USA has been added to the EXCEL database, but not yet to the IDS version (especially marriage dates, characteristics of partners and death dates). The USA was a popular destination among lower-income groups in the second half of the 19th century, especially during the agrarian depression in the 1880s and early 1890s. In the future, more information on the life courses abroad will be added to the database. This addition will allow for new avenues of research that have not been explored yet, such as the socioeconomic success or failure of RPs who moved abroad. Some first results were presented by Koning and Paping (2019) showing a relatively huge upward social mobility of these American migrants.

3.3 STRUCTURE OF THE DATASET

Three cohorts of the rural part of the database (1830, 1850 and 1870) have been converted into the so-called Intermediate Data Structure (IDS; see Alter & Mandemakers, 2014). The IDS has become an increasingly popular format for life course databases (Dribe & Quaranta, 2020; Edvinsson & Engberg, 2020; Jenkinson, Anguita, Paiva, Matsuo, & Matthijs, 2020). Databases in IDS format allow researchers to conduct international and interregional studies with more ease, as it standardises data across databases. The IDS is designed in six tables which key structure can be used in a database management system to connect individuals with each other and with the contexts they are part of on specific moments in time. Until now five out of these six tables have been used for the GIHCD in a Microsoft Access database: INDIVIDUAL,

INDIV_INDIV, CONTEXT, CONTEXT_CONTEXT and METADATA, excluding the table INDIV_CONTEXT. The information included in the five tables is presented below.

The INDIVIDUAL table includes information on personal attributes (e.g., name, occupation) and events (e.g., birth date, marriage date). The basic structure of the table includes a database identifier (*Id_D*) and an individual identifier (*Id_I*) for all included individuals, in our case each RP and the direct relatives that we found and linked to the RPs. Table 5 shows the attributes and events recorded in the GIHCD.

The INDIVIDUAL table counts 90,592 records unevenly belonging to 19,045 unique persons (RPs, parents, spouses and children). Each record consists of a value for a specific type of attributes. Examples of attributes are the RP's first name, the last name and the location of birth. The *Type* Birth_Location can take the *Value* 'Hoogkerk', one of the municipalities. The relatives on which the database provides information is limited to the RPs' parents, their eventual children and marriage partner(s). The information available on these relatives is limited but include sex, birth location, birth date and occupation (only for parents and partners). All occupations are provided with a code number from the Historical International Classification of Occupations (HISCO; van Leeuwen, Maas, & Miles, 2002). Occupational information was collected from civil certificates. For example, the occupations of the parents of an RP are collected from the birth certificate and the occupation of the RP's partner is taken from the marriage certificate. Because more databases also draw on the HISCO codes to make claims about the social position of individuals and their parents, the addition of these codes will allow researchers to easily use the IDS release of this database for comparative research (compare Edvinsson & Engberg, 2020; Mandemakers & Kok, 2020; Vézina & Bournival, 2020).

Finally, the INDIVIDUAL table provides a context identifier (*Value_Id_C*) to connect contextual data to the CONTEXT table. Table 5 outlines some of the aforementioned information that can be found in the INDIVIDUAL table, excluding the time stamps. Three marginal departures from the IDS guidelines stand out. Firstly, in case of the *Type* "Departure to" we have also filled in the name of the location to which RPs moved next to a context identifier. Secondly, we used the value "IntGron_form" for the field *Source* (see Table 6). This value stands for the forms that were used in collecting the data. As mentioned in section 2.3, researchers frequently failed to write down all information on the RPs in the formative years of the database. Source specification was one of the fields that researchers sometimes left empty, which resulted in source specifications with the *Value* "IntGron_form" (3,350 records out of 90,592). Thirdly, in the IDS format, we could not specify the source in case of the *Type* "End_Observation". Contrary to "Start_Observation", for which we automatically could fill in the birth certificate as the only possible source value, the source specification for the *Type* "End_Observation" was sometimes problematic since there are different ends of the observable life course. Observations may end when someone passes away (*Value*: "Death"), when the RP departs out of the register (*Value*: "Departure") to a location in which she or he is not found again, or when the RP is no longer found in registers after being present at the closing of previous registers (*Value*: "End source"). So, for the *Type* "End_Observation" the values for *Source* are empty. See Table 6 for examples of records in the table INDIVIDUAL.

One major departure from the IDS format is the way in which changes in occupation are sometimes denoted. When time stamps were unavailable, the GIHCD indicates changes in occupations by a new occupation *Type* ("Occupation1", "Occupation2", "Occupation3", etc.). Here, "Occupation1" denotes the RP's occupation before and at marriage and, wherever applicable, "Occupation2" and higher denote changes in occupations after marriage. This choice was made because several of the paper forms on the basis of which the database is constructed failed to report the dates on which a new occupation was assigned or reported after marriage. For some occupations this exists on the paper forms but the dates have not been digitised yet. Therefore, it was not possible to indicate the timing of occupation changes, but it was required to use the new developed occupation *Types*. This is something that needs to be adjusted in the future to ensure that the GIHCD data can be combined with data from other IDS databases.

The INDIV_INDIV table records the relationship between the individuals in a database which was in our case restricted to the relations between the RPs and his/her parents, spouse(s) and children. This table uses a second individual identifier (*Id_I_2*) which links the second individual to the first individual (*Id_I_1*) in each row. The field *Relation* contains a value that describes the relationship between both individuals (Alter & Mandemakers, 2014). In Table 7 the structure of the relationships of the first RP of our database is presented, excluding time stamps.

Table 5 *Attributes and events covered in the INDIVIDUAL table*

Attribute or Event	Type name	For RPs	For RPs' Partner(s)	For Parents	For Children
First name	First_name	X			
Last name	Last_name	X			
Sex	Sex	X	X	X	
Date of birth	Birth_date	X	X	X	X
Place of birth	Birth_location	X	X	X	X
Marriage date(s)	Marriage_date	X	X	X	
Marriage location(s)	Marriage_location	X	X		
Date of death	Death_date	X			
Place of death	Death_Location	X			
Location of migration	Departure_to	X			
Occupational titles	Occupation, Occupation1, Occupation2, etc.	X	X	X*	
HISCO code	Occupation_HISCO, HISCO1, etc.	X	X	X	

* *The only occupation of the parents is the one reported on the birth certificates of the RPs.*

Table 6 *Examples of information in the table INDIVIDUAL*

ID	Id_D	Id_I	Source	Type	Value	Value_Id_C
1	IntGron 2015_02	1	Birth certificate	Birth_Date	<time stamp>	
5	IntGron 2015_02	1	Birth certificate	Birth_Location	Hoogkerk	19
2	IntGron 2015_02	1	Death certificate	Death_Date	<time stamp>	
130961	IntGron 2015_02	1	Population register	Departure_To	Bedum	11
132973	IntGron 2015_02	1	Population register	Departure_To	Hoogkerk	19
29161	IntGron 2015_02	1	Death certificate	End_Observation	Death	
4	IntGron 2015_02	1	Birth certificate	First_Name	Goossen	
3	IntGron 2015_02	1	Birth certificate	Last_Name	Aalfs	
77772	IntGron 2015_02	1	Marriage certificate	Marriage_Date	<time stamp>	
82516	IntGron 2015_02	1	Marriage certificate	Marriage_Location	Bedum	11
106307	IntGron 2015_02	1	Marriage certificate	Occupation1	Zonder	
106308	IntGron 2015_02	1	IntGron_form	Occupation2	landbouwer	
6	IntGron 2015_02	1	Birth certificate	Sex	Male	
22681	IntGron 2015_02	1	Birth certificate	Start_Observation	Birth	

In the CONTEXT table descriptive information of the selected municipalities, the city of Groningen and all the localities appearing in the sources as locations of RPs residence, is stored. The attributes (*Type*) are limited to the names, longitudinal centroid, latitudinal centroid, and the type of locality (hamlet, village, town, city, or municipality) and a context identifier (*Id_C*). The spatiotemporal data is drawn from the Dutch Toponyms Spatio-Temporal 1812–2012 database of the HSN (Huijsmans, 2013). Households have not been added, nor families; even though the data from the INDIV_INDIV table allows for some basic family reconstitutions which could include the parents and children of RPs but no other kin, such as RPs' siblings.

The CONTEXT_CONTEXT table embeds the villages, hamlets and towns in the municipalities they are part off. The designation of these localities to their corresponding municipality is also based on the Dutch Toponyms Spatio-Temporal 1812–2012 database. Similar to the INDIV_INDIV table, this table links localities to their municipality by linking two identifiers (*Id_C_1* and *Id_C_2*) and defining the relationship (field *Relation*) between the context layers.

Table 7 *Example of records in the table INDIV_INDIV*

ID	Id_I_1	Id_I_2	Id_D	Source	Relation
1823265	1	3241	IntGron 2015_02	Marriage certificate	Husband
1824949	3241	1	IntGron 2015_02	Marriage certificate	Wife
1794525	1	6308	IntGron 2015_02	Birth certificate	Father
1801495	6308	1	IntGron 2015_02	Birth certificate	Child
1796648	1	8866	IntGron 2015_02	Birth certificate	Father
1801496	8866	1	IntGron 2015_02	Birth certificate	Child
1815102	1	12563	IntGron 2015_02	Birth certificate	Child
1808622	12563	1	IntGron 2015_02	Birth certificate	Father
1818342	1	15803	IntGron 2015_02	Birth certificate	Child
1811862	15803	1	IntGron 2015_02	Birth certificate	Mother

Although the specifications for the IDS (Alter & Mandemakers, 2014) emphasise that the INDIV_CONTEXT table should connect individuals to their temporal and spatial contexts, such as addresses or municipalities, this table has not been filled out so far. The inclusion of information regarding individuals' migrations has been limited to the INDIVIDUAL table by the *Type* "Departure_to". We do not expect this to be an issue for researchers as the only contexts to which RPs can be linked to are limited to hamlets, villages, towns and municipalities and not to lower-level contexts such as households, addresses or families. This means that "Departure_to" captures the full extent of an RPs contextual changes. However, as shown by Alter, Newton, and Oeppen (2020) there are ways to create a context out of a nuclear family that is compatible with IDS, which could also be done for this database. Adding an INDIV_CONTEXT table that connects individuals to families could open up new research avenues in the future.

The METADATA table is the fourth version of the IDS (4.01) with the addition of six new occupation *Types* ("Occupation1"–"Occupation6") which are employed by the GIHCD. When time stamps for all available occupations are traced, these additional *Types* will be replaced by time stamped occupations.

Finally, time stamps follow the latest IDS guidelines. *Date_Type* indicates whether the date is an event that is observed on the event itself, like a birth date on a birth certificate, whether it is reported on a later date, or declared at a point for which an attribute is valid or assigned. *Estimation* indicates whether a date is exact, a middling year between two possible dates, a year, or period with a *Start_Year* and an *End_Year*.

3.4 STRENGTHS AND WEAKNESSES OF THE GIHCD

Presently, for only 39 out of the 3,240 rural cohort members of 1830, 1850 and 1870 no death date or migration outside of the Netherlands has been found. This means that the fate of slightly more than 1% of them is still missing (see tables 3, 4 and the Appendix). The most recent tracings of these RPs suggest that presumably a large part of them will have left the country or died on sea. This makes the GIHCD of excellent quality, at least when it comes to the coverage of the life courses of the persons involved. The extensive research process in the last decades has shown that in particular those RPs with a diverging life course were difficult to follow, for instance those migrating over larger distances and/or remaining unmarried, born outside marriage and those changing names.

Weak points of the GIHCD are: 1. Part of the information is still not digitalised and available on paper only; 2. Due to the changes in the collecting method, some information has not been systematically recorded, as for instance the changes in the precise household structure of the cohort members; 3. Although there is a lot of new information collected on the urban cohorts (4 * 240 = 960 RPs) and the 1811 rural cohorts (1,080 RPs), this has not yet been systematically included in the digital database; 4. The limited embedding in an organisational structure since the last 25 years, makes the database the sole responsibility of one person.

4 OUTPUT

As previously stated, the impact of the GIHCD has been limited, mainly because it has only been used by economic and social historians of the University of Groningen. Consequently, the studies that have been conducted so far refer to debates in historical demography and economic and social history. We will briefly address them in this section.

The first publication that made use of GIHCD data was the volume on the village of Hoogkerk between 1770 and 1914 (Kooij, 1993). Hoogkerk was selected for a pilot study in which various relevant questions and methods could be tested, especially because of its spectacular transformation from an agrarian village to an industrialised area around 1900. The book ambitiously aimed at integrating the various domains that explained social developments in the village through a lens that combined economic, political, social, cultural, religious and demographic elements. With the help of the GIHCD, the demographic and socioeconomic developments could be disentangled to some extent. Although the RP sample for the four cohorts from Hoogkerk was small ($n = 480$), it was used by Marcel Clement (1993) to analyse demographic developments in Hoogkerk on a micro level. He studied (child) mortality, nuptiality, migration and social mobility. Due to high levels of child mortality and migration, only a small amount of the RPs delivered substantial data in Hoogkerk. But it showed how mobile 19th-century Dutch people were, as almost every RP left Hoogkerk at least once.

Within the field of historical demography, the interest in individual agency grew in the last decades of the 20th century. How and why families and individuals make choices or were forced to do so in order to improve their socioeconomic position (Engelen, Knotter, Kok, & Paping, 2004). This turn towards agency stimulated the analysis of short- and long-term decisions of households, families and persons, taking into account the wider historical context. However, quantitative data that only present static information on peoples' location or occupation, and do not show how these changed over time, is not suited to easily assess the motives behind the choices. As a possible solution, Paping (1999, pp. 18–19) used the GIHCD to explain how and why different socioeconomic classes made different short- and long-term choices regarding their employment, the employment of their children and their migration patterns. The study showed that married couples often migrated within the first years after marriage. Afterwards their inclination to migrate diminished rapidly, suggesting a rise in local social embedment over time. Furthermore, it was shown using the information on moments of leaving and juvenile occupations in the GIHCD that relatively many lower-income (unskilled) families opted for short-term strategies, like sending children away as live-in servants, which proved not very beneficial in the long run (Paping, 2004a, p. 188). For lower-income families, the employment of children as servants from the age of about 14 was one of the few options available to avoid the costs of having largely unemployed older children at home. Higher income groups, such as large farmers and the self-employed in industry and services, could afford to keep their children at home. For every social group, those children remaining at home proved to acquire on average a much better position in the long run than those becoming servants. That unskilled labourers responded to real wage increases at the end of the 19th century by increasingly keeping their children at home, suggests that they saw this choice indeed as a positive long-term strategy (see also Paping, 2017). However, changing views on education in the late 19th century might also have played a role (Kooij, 2004, p. 196).

In another publication, Paping (2004b) used part of the cohort database to examine the strategies families employed with regards to the labour of their family members. With the addition of financial microdata from various Nieuw Scheemda farmers' bookkeeping, Paping focused on the group of unskilled farm labourers in Groningen. The financial situation of unskilled worker families proved to be very volatile in the second half of the 19th century. With the exception of the male household heads, the rest of the family members — both wives and children — heavily depended on casual or seasonal labour. Adolescent sons and daughters usually left the household rather soon to become live-in servants. Information on the dates of marriage and the birth of the first child in the database showed that forced marriages due to pregnancies were common among unskilled worker couples. Such a situation restricted the opportunity for unskilled labourer families to make long-term decisions, as they would be forced to make short-term choices in response to an unplanned pregnancy.

About a decade after the pilot study on Hoogkerk, Kooij (2004) repeated some of the principles of the Integral History Project in a theoretical essay. One of these principles, the necessity to integrate the various domains that influenced peoples' lives had not changed. Kooij (2004, pp. 194) distinguished the economic, demographic, social, cultural, political and religious domains as central in understanding the pressures on individual and household decisions.

To analyse the relationship between two of these domains, the cultural/religious domain and the socioeconomic domain, Collenteur and Paping (2004) used rural cohort data from the GIHCD and tried to disentangle the connection between cultural and economic effects on peoples' decision to marry in a comparative contribution. They compared differences in marriage patterns between three Russian regions and two Dutch regions. The Russian regions consisted of the geographically farthest to the east, Tambov (located about 410 kilometers southwest of Moscow), the farthest to the west, Olonets (located about 180 kilometers northwest of St. Petersburg) and Yaroslavl (located about 250 kilometers northwest of Moscow). The Dutch regions were Groningen and North-Brabant. The data showed that, on average, rural Russians married at an earlier age compared to people from the Groningen urban and rural regions, but that differences between the Russian regions were huge. Furthermore, men and women in Groningen showed large variations in marriage age, as it was not uncommon to marry around the age of 20, although it was not rare for couples to wait until their 30s. Collenteur and Paping (2004) concluded that marriage patterns were likely driven largely by socioeconomic factors in Groningen, and RPs had a lot of agency with respect to their marriage age, at least in the absence of unplanned pregnancies. This was absolutely not the case in the Tambov region, the farthest east of the Russian regions studied. In Tambov, young adults usually married before their 20s with very small variation in ages at marriage, indicating an important role for cultural factors like traditions and norms influencing marriage decisions, leaving very limited room for individual agency. Although the differences between the Russian regions studied in this research proved to be large, the data suggest that the further east someone went, the lower the average age of marriage and the smaller the variation in age at marriage would be. In this way, Collenteur and Paping's research provide some support for Hajnal's (1983) hypothesis on marriage patterns and the so-called Hajnal line.

Kooij (2011) used the 1830 and 1870 cohorts of Beerta and Winschoten to analyse similarities and differences between a more urban-oriented region (Winschoten) and a completely rural region (Beerta) in the province of Groningen. He shortly compared various social groups on the basis of their life chances and mobility, and did not find major differences in life expectancy and geographic mobility between urban and rural environments. However, he concluded that for the 1870 cohort long distance-migration became more common, both within the Netherlands and outside of it. Furthermore, he showed that the agrarian depression of the late-19th century gave rise to chain migration towards the USA.

Finally, Paping and Pawlowski (2018) compared intergenerational occupational social mobility of rural-to-urban migrants with other groups, such as rural stayers, in the Groningen province. Their study used two databases. First, they employed the huge AlleGroningers database with summaries of all 234,000 marriages concluded in the province of Groningen between 1811 and 1934, out of which those of persons born in the Groningen clay area (see figure 1) were selected (N = 121,000). Second, the much smaller GIHCD was used, which allowed studying the relation between social mobility and rural-urban migration much more in depth, taking into account migration to more distant Dutch cities, as well as for instance migration taking place after marriage.

Rural males moving to an urban environment showed both much more upward and downward social mobility, compared to rural stayers or rural to rural migrants, who showed high social immobility. Furthermore, the findings reveal that urban *pull factors* were primarily responsible for the rural to urban migration, especially the better employment opportunities in the cities. This view contradicts a common explanation that rural-urban migration was mainly stimulated by bad rural circumstances forcing rural poor to move. Just the opposite took place: especially those descending from the somewhat higher social classes were much more inclined to move to the cities, as they had more useful skills than unskilled labourers for an urban environment. Although also experiencing relatively higher chances on downward mobility, migrants moving to cities showed, on average, even more upward social mobility than rural stayers in almost every social group. The GIHCD also made it possible to look at short stayers. According to some research (see Puschmann, 2015), positive results for rural to urban migration are biased upward because they do not account for migrants that return after a short stay in the city. In this view, a return to the rural environment likely represented a failed move to the city. Paping and Pawlowski (2018), however, show that even a short stay in an urban environment led on average to lasting positive effects on the social position for most socioeconomic groups.

5 CONCLUDING REMARKS

The success of the so-called cohort analysis in the Integral History Project was restricted by the combination of ambitions that were too high, some naïve decisions being made and the lack of structural funding,

in addition to the consequences of enormous and rapid technical changes taking place since the 1980s. Until 1995, the quality of the database was too low, as a consequence the research output was not very impressive. Later on, the quality improved and several meaningful scientific publications were based on it. However, the use of the database remained difficult due to the struggle to regularly update it technically, and to incorporate new digitally available sources, without proper funding.

Despite all these drawbacks, in one respect the quality of the GIHCD is extremely high. In contrast with other databases only less than 2% of persons involved could not be traced during their whole life course, making it very representative as it also includes the most extreme and rare cases that in other databases are often missing. This is a major reason to try to make the database better available in the future, both in an IDS structure and as an EXCEL spreadsheet.

The relatively small size remains problematic, especially in this time of digitalisation. It seems rather odd to study only a sample of the population, while — at least for part of the less complex questions — you can just as well use the enormous databases now available in the Netherlands. After all, as the quality of software-based record linkages is still improving, the future might not be for samples as the GIHCD, but for enormous databases connecting events of all persons in a geographical region, at least for the 19th and 20th century (see for northwest Groningen: Paping & Schansker, 2013, 2014), and maybe even for an earlier period. However, as long as not all the relevant sources have been digitalised in a way that makes linking possible, and as long as digitalised linkage procedures are still not perfect, there is a need for high quality specific databases like the Groningen Integral History Cohort Database.

REFERENCES

- Akolzina, M., Dyatchkov, V., Kanitshev, V., Kontchakov, R., Mizis, Y., & Morozova, E. (2004). A comparison of cohort analysis and other methods of demographic microanalysis used in studying the Tambov region. In P. Kooij & R. Paping (Eds.), *Where the twain meet again: New results of the Dutch-Russian project on regional development 1780–1917* (pp. 45–90). Groningen/Wageningen: NAHI. Retrieved from <https://ugp.rug.nl/ha/article/view/2103>
- Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies*, 1, 1–26. doi: 10.51964/hlcs9290
- Alter, G., Newton, G., & Oeppen, J. (2020). Re-introducing the Cambridge Group family reconstitutions. *Historical Life Course Studies*, 9, 24–48. doi: 10.51964/hlcs9311
- Clement, M. (1993). Demografisch gedrag, leefsituatie en mobiliteit. Een analyse van vier generaties. In P. Kooij (Ed.), *Dorp naast een stad: Hoogkerk 1770–1914* (pp. 160–201). Assen: Van Gorcum.
- Collenteur, G., & Paping, R. (2004). Age at first marriage in eighteenth and nineteenth-century Russia and the Netherlands: Tradition or economic and social circumstances? In P. Kooij & R. Paping (Eds.), *Where the twain meet again: New results of the Dutch-Russian project on regional development 1780–1917* (pp. 147–167). Groningen/Wageningen: NAHI. Retrieved from <https://ugp.rug.nl/ha/article/view/2103>
- Commissie voor de statistieke beschrijving der provincie Groningen (Eds.). (1870). *Bijdragen tot de kennis van den tegenwoordigen staat der provincie Groningen* (Vijfde deel: Landbouw-statistiek). Groningen: Hoitsema.
- Dribe, M., & Quaranta, L. (2020). The Scanian Economic-Demographic Database (SEDD). *Historical Life Course Studies*, 9, 158–172. doi: 10.51964/hlcs9302
- Duijvendak, M. G. J., & Blijham G. J. (1994). Groninger provinciale financiën in Nederlands perspectief 1824–1910. *NEHA-Jaarboek voor Economische, Bedrijfs- en Techniekgeschiedenis*, 57, 206–248. Retrieved from <https://webstore.iisg.nl/neha/13805517-1994-001.pdf>
- Dyatchkov, V., Kanitshev, V., Mizis, Y., Orlova, V., Protasov, L., & Protasov, S. (1998). Cohort analysis of Malye Pupky's population: Some preliminary results. In P. Kooij (Ed.), *Where the twain meet. Dutch and Russian regional development in a comparative perspective 1800–1917* (pp. 141–154). Groningen: NAHI.
- Edvinsson, S., & Engberg, E. (2020). A database for the future. Major contributions from 47 years of database development and research at the Demographic Data Base. *Historical Life Course Studies*, 9, 173–196. doi: 10.51964/hlcs9305
- Engelen, T., Knotter, A., Kok, J., & Paping, R. (2004). Labor strategies of families: An introduction. *The History of the Family*, 9(2), 123–135. doi: 10.1016/j.hisfam.2004.01.001

- Golubeva, S. (1998). Age and patterns of marriage of Russian farmers in the Yaroslavl region. In P. Kooij (Ed.), *Where the twain meet. Dutch and Russian regional development in a comparative perspective 1800–1917* (pp. 169–174). Groningen: NAHI.
- Hajnal, J. (1983). Two kinds of pre-industrial household formation system. In R. Wall, J. Robin & P. Laslett (Eds.), *Family forms in historic Europe* (pp. 65–104). Cambridge: Cambridge University Press.
- Hoogeboom, B. (2013). *Van de Ommelanden naar de Martinistad. Migratie naar de stad als strategie om te ontsnappen aan relatieve rurale achteruitgang* (Unpublished bachelor thesis). University of Groningen, Groningen.
- Huijsmans, D. P. (2013). *IISG-LINKS Historische Nederlandse Toponiemen Sption-Temporeel 1812–2012. Release 2013.2*. Retrieved from <https://iisg.amsterdam/en/hsn/data/place-names>
- Integral History Project Groningen. (n.d.). Retrieved from <https://ehps-net.eu/databases/integral-history-project-groningen>
- Jenkinson, S., Anguita, F., Paiva, D., Matsuo, H., & Matthijs, K. (2020). The 2020 IDS release of the Antwerp COR*-database. Evaluation, development and transformation of a pre-existing database. *Historical Life Course Studies*, 9, 197–217. doi: [10.51964/hlcs9301](https://doi.org/10.51964/hlcs9301)
- Klokkenburg, L. (2009). *Afgescheiden en afgezonderd? Het gedrag van de afgescheidenen in de Groninger klei gedurende de negentiende eeuw* (Unpublished master thesis). University of Groningen, Groningen.
- Koning, P. J. (2019). *Gouden bergen of de Groninger klei? Een case study over sociale mobiliteit van migranten uit de Groningse kleigebieden naar de Verenigde Staten tussen 1850 en 1940* (Unpublished master thesis). University of Groningen, Groningen.
- Koning, P. J., & Paping, R. (2019, November). Economic Refugees' from the northern Dutch countryside? Social origin and socio-economic success of Dutch migrants to the USA in the second half of the 19th century. Presentation at the *10th Day of the Historical Demography*, Nijmegen.
- Kooij, P. (1987). *Groningen 1870–1914. Sociale verandering en economische ontwikkeling in een regionaal centrum*. Assen/Maastricht: Van Gorcum.
- Kooij, P. (Ed.). (1993). *Dorp naast een stad: Hoogkerk 1770–1914*. Assen: Van Gorcum.
- Kooij, P. (1993a). Inleiding. In P. Kooij (Ed.), *Dorp naast een stad: Hoogkerk 1770–1914* (pp. 1–5). Assen: Van Gorcum.
- Kooij, P. (1993b). Bevolking: Huishoudens, gezinnen en sociale stratificatie. In P. Kooij (Ed.), *Dorp naast een stad: Hoogkerk 1770–1914* (pp. 130–159). Assen: Van Gorcum.
- Kooij, P. (Ed.). (1998). *Where the twain meet. Dutch and Russian regional development in a comparative perspective 1800–1917*. Groningen: NAHI.
- Kooij, P. (2004). Demographic development in the context of integral history, 1800–1917. In P. Kooij & R. Paping (Eds.), *Where the twain meet again: New results of the Dutch-Russian project on regional development 1780–1917* (pp. 191–198). Groningen/Wageningen: NAHI. Retrieved from <https://ugp.rug.nl/ha/article/view/2103>
- Kooij, P. (2011). Uit de klei getrokken. De levensloop van migranten uit de Oost-Groninger geboortecohorten Beerta en Winschoten 1830 en 1870. In T. Engelen, O. Boonstra & A. Janssens (Eds.), *Levenslopen in transformatie: Liber Amicorum bij het afscheid van prof. dr. Paul M. M. Klep* (pp. 279–290). Nijmegen: Valkhof Pers.
- Kooij, P., & Paping, R. (Eds.). (2004). *Where the twain meet again: New results of the Dutch-Russian project on regional development 1780–1917*. Groningen/Wageningen: NAHI. Retrieved from <https://ugp.rug.nl/ha/article/view/2103>
- Kooij, P., & Sleebe, V. (1991). A small village in a changing world; Integral history at a local level. *Economic and Social History in the Netherlands*, 3, 19–36.
- Mandemakers, K. (2000). The Netherlands. Historical Sample of the Netherlands. In: P. Kelly Hall, R. McCaa & G. Thorvaldsen (Eds.), *Handbook of International Historical Microdata for Population Research* (pp. 149–177). Minneapolis: Minnesota Population Center.
- Mandemakers, K., Bloothoof, G., & Laan, F. (forthcoming). LINKS. The LINKing System for historical family reconstruction in the Netherlands. *Historical Life Course Studies*.
- Mandemakers, K., & Kok, J. (2020). Dutch Lives. The Historical Sample of the Netherlands (1987–): Development and Research. *Historical Life Course Studies*, 9, 69–113. doi: [10.51964/hlcs9298](https://doi.org/10.51964/hlcs9298)
- Paiva, D., Anguita, F., & Mandemakers, K. (2020). Linking the Historical Sample of the Netherlands with the USA censuses, 1850–1940. *Historical Life Course Studies*, 9, 1–23. doi: [10.51964/hlcs9312](https://doi.org/10.51964/hlcs9312)
- Paping, R. (1995). *Voor een handvol stuivers. Werken, verdienen en besteden: de levensstandaard van boeren, arbeiders en middenstanders op de Groninger klei, 1770–1860* (Doctoral dissertation). Groningen: NAHI. Retrieved from <https://pure.rug.nl/ws/portalfiles/portal/14976871/thesispaping1995.pdf>

- Paping, R. (1999). Gezinnen en cohorten: Arbeidsstrategieën in een marktgerichte agrarische economie: De Groningse kleigebieden 1830–1920. In J. Kok, A. Knotter, R. Paping & E. Vanhaute (Eds.), *Levensloop en levenslot. Arbeidsstrategieën van gezinnen in de negentiende en twintigste eeuw* (pp. 17–88). Groningen/Wageningen: NAHI. Retrieved from https://www.researchgate.net/profile/Jan-Kok-4/publication/333401946_Levensloop_en_levenslot/links/5cebd797a6fdcce250aa63ca/Levensloop-en-levenslot.pdf
- Paping, R. (2004a). Family strategies concerning migration and occupations of children in a market-oriented agricultural economy. *The History of the Family*, 9(2), 159–191. doi: 10.1016/j.hisfam.2004.01.003
- Paping, R. (2004b). Family strategies, wage labour and the family life cycle in the Groningen countryside, c. 1850–1910. In P. Kooij & R. Paping (Eds.), *Where the twain meet again: New results of the Dutch-Russian project on regional development 1780–1917* (pp. 271–291). Groningen/Wageningen: NAHI. Retrieved from <https://ugp.rug.nl/ha/article/view/2103>
- Paping, R. (2009). Gender and the intergenerational transfer of property and social position in the 18th and early 19th century northern Dutch countryside. In M. Durães, A. Fauve-Chamoux, L. Ferrer & J. Kok (Eds.), *The transmission of well-being: Gendered marriage strategies and inheritance systems in Europe (17th–20th centuries)* (pp. 291–313). Bern: Peter Lang. Retrieved from https://www.researchgate.net/publication/260081918_Gender_and_the_intergenerational_transfer_of_property_and_social_position_in_the_18th_and_early_19th_century_northern_Dutch_countryside
- Paping, R. (2010). Taxes, property size, occupations and social structure: The case of the 18th and 19th century northern Dutch countryside. *Belgisch Tijdschrift voor Nieuwste Geschiedenis*, 40(1), 215–248. Retrieved from <https://www.journalbelgianhistory.be/nl/journal/belgisch-tijdschrift-voor-nieuwste-geschiedenis-2010-1-2/taxes-property-size-occupations>
- Paping, R. (2017). Dutch live-in farm servants in the long 19th century: The decline of the life-cycle service system for the rural lower class. In J. Whittle (Ed.), *Servants in rural Europe. 1400–1900* (pp. 203–226). Woodbridge: Boydell & Brewer.
- Paping, R. (2018). The 'dynamics' of household structure and size in the Northern Dutch countryside around 1840. In P. Puschmann & T. Riswick (Eds.), *Building bridges. Scholars, history and historical demography. A Festschrift in honor of professor Theo Engelen* (pp. 362–384). Nijmegen: Valkhof Pers. Retrieved from https://www.ru.nl/publish/pages/800120/building_bridges_open_access_pdf.pdf
- Paping, R., & Collenteur, G. (1998). The economic development of the clay soil area of Groningen 1770–1910: Income and socio-economic groups. In: P. Kooij (Ed.), *Where the twain meet. Dutch and Russian regional development in a comparative perspective 1800–1917* (pp. 35–50). Groningen: NAHI.
- Paping, R., & Collenteur, G. (2004). Population growth and social structure in a market-oriented agricultural economy in The Netherlands 1750–1820. *Obradoiro de Historia Moderna*, 13, 75–99. Retrieved from http://dspace.usc.es/bitstream/10347/3936/1/pg_076-101_obradoiro13.pdf
- Paping, R., & Pawlowski, J. (2018). Success or failure in the city? Social mobility and rural-urban migration in nineteenth- and early-twentieth-century Groningen, the Netherlands. *Historical Life Course Studies*, 6, 69–94. doi: 10.51964/hlcs9329
- Paping, R., & Schansker, G. (2013). The reproduction of the rural labour class: Fertility, nuptiality and life chances. Paper presented at the 6th Day of the Historical Demography. Retrieved from <https://www.researchgate.net/publication/282249917>
- Paping, R., & Schansker, G. (2014). De reproductie van de rurale arbeidersklasse in achttiende- en negentiende-eeuws Groningen: Vruchtbaarheid, nuptialiteit en overlevingskansen. In I. Devos, K. Matthijs & B. Van de Putte (Eds.), *Kwetsbare groepen in/en historische demografie* (pp. 71–98). Leuven/Den Haag: Acco.
- Puschmann, P. (2015). *Social inclusion and exclusion of urban in-migrants in northwestern European port cities. Antwerp, Rotterdam & Stockholm, ca. 1850–1930* (Doctoral dissertation). Retrieved from https://limo.libis.be/primo-explore/fulldisplay?docid=LIRIAS1871846&context=L&vid=Lirias&search_scope=Lirias&tab=default_tab&lang=en_US&fromSitemap=1
- Schrover, M. (1998). Demographic behaviour in North Brabant in the nineteenth century. In P. Kooij (Ed.), *Where the twain meet. Dutch and Russian regional development in a comparative perspective 1800–1917* (pp. 201–222). Groningen: NAHI.
- Shustrova, I., & Sinitsyna, E. (2004). Demographic behaviour in the Yaroslavl loamy area. The results of cohort analysis for two typical rural villages. In P. Kooij & R. Paping (Eds.), *Where the twain meet again: New results of the Dutch-Russian project on regional development 1780–1917* (pp. 7–18). Groningen/Wageningen: NAHI. Retrieved from <https://ugp.rug.nl/ha/article/view/2103>
- Sinitsyna, E. (1998). Some results of cohort analysis in the Yaroslavl region. In P. Kooij (Ed.), *Where the twain meet. Dutch and Russian regional development in a comparative perspective 1800–1917* (pp. 155–168). Groningen: NAHI.

- Sleebe, V. (1993). Openbaar en privé. In P. Kooij (Ed.), *Dorp naast een stad: Hoogkerk 1770–1914* (pp. 295–335). Assen: Van Gorcum.
- Trienekens, G. (1987). Theoretische en methodologische aspecten van de lokale en regionale geschiedschrijving. In F. van Besouw, P. den Boer, F. W. N. Hugenholtz & Th. van Tijn (Eds.), *Balans en perspectief. Visies op de geschiedwetenschap in Nederland* (pp. 167–188). Groningen: Wolters-Noordhoff/Forsten.
- Trienekens, G. (1993). Integrale geschiedenis in wording: Aarle-Rixtel en Wanroij in de negentiende en het begin van de twintigste eeuw. In J. A. van Oudheusden & G. Trienekens (Eds.), *Een pront wijf, een mager paard en een zoon op het seminarie: Aanzetten tot een integrale geschiedenis van oostelijk Noord-Brabant* (pp. 211–313). 's-Hertogenbosch: Stichting Brabantse Regionale Geschiedoefening.
- Trienekens, G. (2004). Characteristics of households in the eastern part of North Brabant 1810–1920. In P. Kooij & R. Paping (Eds.), *Where the twain meet again: New results of the Dutch-Russian project on regional development 1780–1917* (pp. 25–44). Groningen/Wageningen: NAHI. Retrieved from <https://ugp.rug.nl/ha/article/view/2103>
- van Leeuwen, M. H. D., Maas, I., & Miles, A. (2002). *HISCO. Historical International Standard Classification of Occupations*. Leuven: Leuven University Press.
- van Oudheusden, J. A., & Trienekens, G. (Eds.) (1993). *Een pront wijf, een mager paard en een zoon op het seminarie: aanzetten tot een integrale geschiedenis van oostelijk Noord-Brabant*. 's-Hertogenbosch: Stichting Brabantse Regionale Geschiedoefening.
- Vézina, H., & Bournival, J. S. (2020). An overview of the BALSAC population database. Past developments, current state and future prospects. *Historical Life Course Studies*, 9, 114–129. doi: [10.51964/hlcs9299](https://doi.org/10.51964/hlcs9299)
- Voerman, J. F. (2001). *Verstedelijking en migratie in het Oost-Groningse veengebied 1800–1940* (Doctoral dissertation). Assen: Van Gorcum.

APPENDIX

Table A1 *Life course characteristics of the 1811 cohort*

1811	Total	Lost	Emigrated permanently	Died before 20 in the Netherlands	Died after 20 in the Netherlands	Marriage year known
City of Groningen	240	16	9	86	129	110
Appingedam	120	5	2	39	74	61
Beerta	120	2	5	37	76	66
Bedum	120	0	3	36	81	66
Hoogkerk	120	2	2	38	78	60
Leens	120	3	12	33	72	64
Stedum	120	4	6	34	76	66
Uithuizen	120	5	8	32	75	73
Winschoten	120	1	4	26	89	76
Zuidhorn	120	1	3	27	89	74
Total rural	1,080	23	45	302	710	606

Table A2 *Life course characteristics of the 1830 cohort*

1830	Total	Lost	Emigrated permanently	Died before 20 in the Netherlands	Died after 20 in the Netherlands	Marriage year known
City of Groningen	240	9	4	94	133	97
Appingedam	120	4	5	35	76	64
Beerta	120	3	4	20	93	83
Bedum	120	1	8	33	78	70
Hoogkerk	120	1	3	28	88	74
Leens	120	4	13	42	61	58
Stedum	120	1	15	33	71	72
Uithuizen	120	0	28	27	65	69
Winschoten	120	4	4	30	82	67
Zuidhorn	120	1	6	29	84	74
Total rural	1,080	19	86	277	698	631

Table A3 *Life course characteristics of the 1850 cohort*

1850	Total	Lost	Emigrated permanently	Died before 20 in the Netherlands	Died after 20 in the Netherlands	Marriage year known
City of Groningen	240	6	0	90	139	116
Appingedam	120	2	5	40	73	65
Beerta	120	1	7	30	82	71
Bedum	120	1	3	30	86	75
Hoogkerk	120	1	10	36	73	61
Leens	120	1	24	34	61	63
Stedum	120	1	21	43	55	62
Uithuizen	120	1	31	36	52	59
Winschoten	120	4	7	34	75	67
Zuidhorn	120	1	16	30	73	80
Total rural	1,080	13	124	313	630	603

Table A4 *Life course characteristics of the 1870 cohort*

1870	Total	Lost	Emigrated permanently	Died before 20 in the Netherlands	Died after 20 in the Netherlands	Marriage year known
City of Groningen	240	0	12	99	129	116
Appingedam	120	0	11	45	64	61
Beerta	120	2	25	40	53	67
Bedum	120	0	16	43	61	65
Hoogkerk	120	1	7	46	66	59
Leens	120	1	21	37	61	67
Stedum	120	0	39	44	37	61
Uithuizen	120	2	42	31	45	71
Winschoten	120	1	7	39	73	67
Zuidhorn	120	0	12	37	71	67
Total rural	1,080	7	180	362	531	585

HISTORICAL LIFE COURSE STUDIES
VOLUME 13 (2023), published 11-07-2023

GENEVA

An Urban Sociodemographic Database

Michel Oris	Institute of Demography and Socioeconomics and Centre LIVES, University of Geneva & Spanish Research Council, Madrid
Oliver Perroux	Collège de Saussure, Geneva
Grazyna Ryczkowska	Collège de l'Union, Prilly
Reto Schumacher	Cantonal Statistical Office, Vaud
Adrien Remund	Population Research Centre, Faculty of Spatial Sciences, University of Groningen
Gilbert Ritschard	Institute of Demography and Socioeconomics and Centre LIVES, University of Geneva

ABSTRACT

The Geneva databases are a data resource covering the period 1800–1880 for the city of Geneva, and occasionally the canton of Geneva. The research team adopted an alphabetical sampling approach, collecting data on individuals whose surname begins with the letter B. The individuals and households belonging to this sample in six population censuses between 1816 and 1843 were digitised and linked. A second database collected marriage and divorce records for the period 1800–1880. A third collection of data included residence permits. All these sources were used for a massive reconstitution of families. This article presents the sources, the linking methods, the typologies used to code places and occupations, to study household structures and forms of solitude. Combined with qualitative information extracted from the archives of public administrations and the National Protestant Church, as well as from newspapers, these databases were used to study the transformation of a medium-sized European city, sociopolitical tensions embedded in demographic and social structures, and the impact of the immigrants who made the 'Calvinist Rome' a religiously mixed city.

Keywords: Geneva, Historical demography, Censuses, Marriages, Divorces, Residence permits

DOI article: <https://doi.org/10.51964/hlcs15621>

© 2023, Oris, Perroux, Ryczkowska, Schumacher, Remund, Ritschard

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

In Bardet and Dupâquier's *Histoire des Populations de l'Europe* (1998), Geneva was the third most cited city, after London and Paris. This position may come as a surprise given its much smaller population. Disproportionate interest for Geneva stems from the second half of the 16th century, when the city became the "Calvinist Rome", a land of refuge, a religious and intellectual beacon (Zemon Davis, 2015). Waves of refugees changed the demography of Geneva (Perrenoud, 1979). Moreover, the local data sources provoked the interest of Louis Henry, the founding father of historical demography. Before applying his method of family reconstitution on a rural population (Crulai, see Henry & Gauthier, 1958), he tested this method on the old Genevan families (Henry, 1956). Following in his footsteps, Alfred Perrenoud (1979) convincingly extended this pioneering work and carried out one of the first major demographic studies of an urban population in early modern time. Exploiting a system of civil registration created by Jean Calvin himself, he identified the existence of large social gradients in mortality already in a preindustrial context (Bengtsson & van Poppel, 2011; Perrenoud, 1975) and documented the pioneering process of fertility decline observed in Geneva (Perrenoud, 1988).

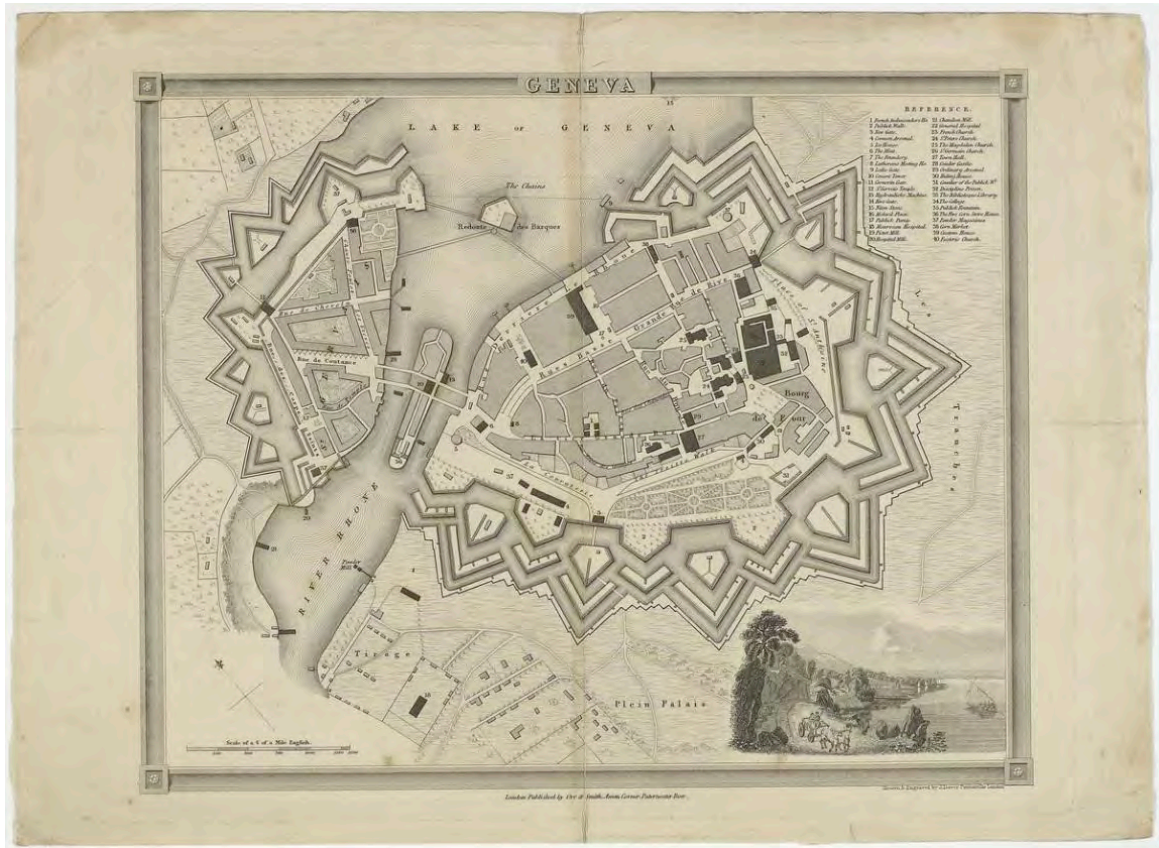
One of our ambitions was to extend the work of Perrenoud, starting where he stopped, that is in the early 19th century. Between 1816 and 1846, Geneva underwent a profound transition. The old republic joined the Swiss Confederation and became a canton composed not only of the Calvinist city but also of surrounding rural municipalities mostly populated by Catholics, thus opening a history of religious cohabitation which was not exempt from social and political tensions. In the meantime, Geneva's population experienced a modest but steady growth. In 1798, the city counted 21,327 inhabitants within its walls, against 31,200 in 1850. This is not an impressive demographic growth in the 19th century European context, which is a consequence of the late industrialization that did not reach the city until late in the century. The dominant sector in the city's economy was the watch-and-clock industry, known locally as the "Fabrique". Far from being organized in factories, this protoindustrial system engaged jewelers and a large number of watchmakers in a vast network of craftsmen. Almost all of this traditional production was exported to international markets, which made both this industry and the city very sensitive to economic fluctuations and political tensions across the European continent. In 1846, the government of the canton, run by an old conservative elite of allied patrician families, was overthrown by a revolution led by the "Radicals", a liberal and pro-democratic movement who recruited most of its supporters among the watchmakers (Perroux, 2006). The new leaders organized the destruction of the old Vauban-style fortifications, which corseted the city and promoted a new phase of urban, economic and social development (see Map 1).

Studying Geneva in the first half of the 19th century means exploring one of the transitions from the *Ancien Régime* to modern economic growth. Such a path is quite different from those followed by the "mushroom towns", cradles of the industrial revolution, or the road taken by larger cities, pillars of the formation of modern state. These industrial cities and metropolises were responsible for most of the 19th century urban expansion and attracted most attention among historical demographers (Ramiro Fariñas & Oris, 2016). The path taken by Geneva, characterized by late and moderate industrialization, a politicized group of craftsmen, tensions between the old political order and aspirations to democracy, as well as between ancient rights of bourgeoisie and modern national citizenship, between locals and newcomers, illustrates the fate of many small- and medium-sized European cities (see Hatt-Diener, 2004; Lorenceau, 2001; Prost, 2011; Reher, 1990; Sewell, 1985), which are still today an important component of the European urban system.

2 SOURCES AND SAMPLES

The reconstruction of an urban population is necessarily a challenge, mainly because of the abundance of archival material and high individual mobility. Several data sources were digitized and linked at the individual and household levels. They are shown in Figure 1, with links that will be explained in the next section.

Map 1 Geneva ± 1829 (John Howe)



Source: Centre genevois d'iconographie.

The first and most important data sources were the population censuses carried out by the Geneva authorities. Between 2003 and 2005, a Swiss National Science Foundation Research Project supported a data extraction from six censuses: 1816, 1822, 1828, 1831, 1837, and 1843. This six-year regular periodicity was relatively exceptional. The exception, in 1831, only three years after the 1828 census, was justified by the fear that provoked the cholera pandemic closing in on Geneva; the authorities wanted to investigate the state of housing and the concentration of the population as a matter of urgency.

The organization of these censuses was based in part on the tight control of the city by the Protestant Church, which had each block of houses supervised by a *dizenier* ('pastor' or 'elder'). Thus, neighborhood by neighborhood, street by street, house by house, each household, the basic unit of the survey, was enumerated, as well as each person in each household. For each individual, the last name, first name(s), marital status, places of birth and origin, residence permits number for non-Genevans, age or date of birth, religion and occupation were recorded, as well as the household address (street and house number). The censuses were carried out within a few days by 80 to 100 agents in 1822 and 1828, and up to 500 in 1843. However, their coverage might have been affected by lack of consensus about whether de jure or de facto populations should be counted (Schumacher 2010, pp. 174–178). Another issue is related to the disappearance of one of the 1831 census registers (see the impact on the sample and the linkages in Table 1).

During the digitalization of these sources, research assistants used an alphabetical sampling approach which ensured the representativeness of the sample. We selected all individuals whose surname began with the letter B. The choice of this letter follows the suggestion of Jacques Dupâquier (1984, p. 115) and Jean-Pierre Bardet (1983), who observed that B is not associated to a specific occupation or social status, nor to an ethnic or linguistic group (that holds for French names as well as Italian and German patronyms). In Geneva, if we add up the city's population over the six censuses between 1816 and 1843, we reach a total of just over 155,000 person-census observations. The B sample we collected comprises 18,976 individual notices, or just over 12% of the total. This proportion remains stable over time and across the various religious and socioeconomic groups. This is the sample that we used in most of our analyses (see below). Our database also includes 16,614 other people, since we sampled all households with at least one member with a last name starting with a B, and digitized all the

individuals within those households. Including all members of a household was essential to be able to study household size and structures.

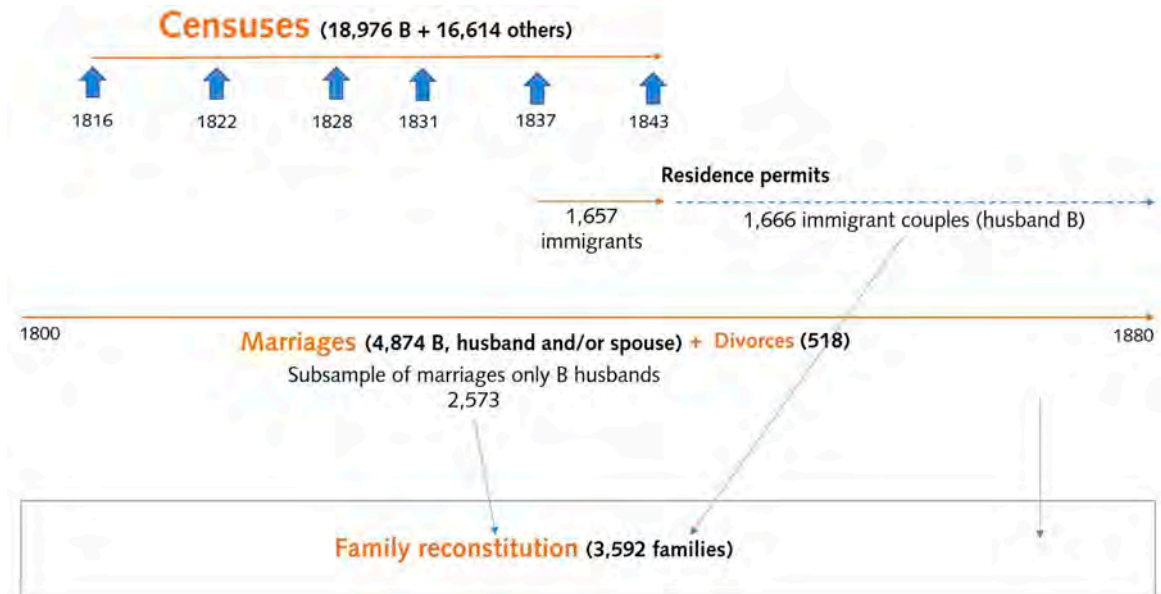
The civil registration was the second most important data source, especially the marriage certificates. Originally, master students working on a practical course of research method initiated the digitalization of those certificates for the period 1800–1840. Later, for her master (2003), then PhD thesis (2013), Grazyna Ryczkowska collected all marriage certificates from 1800 to 1880 of couples among which either the bride or the groom had a surname beginning with the letter B. She did this first for the city (n=4,874) then for the whole canton (n=8,506).

The marriage ceremony was an important source of tensions between Protestants and Catholics. In 1816, for the first time since the Reformation, Catholics and Protestants officially had the same rights in the city (and canton) and were thus subject to the same duties. In practice however, the former conservative Protestant elite regained power and tried to preserve their religious and political domination over the new canton. In this context, the Napoleonic Civil Code, introduced under the Empire which remained in force in Geneva throughout the 19th century, was compatible with the old Calvinist edicts, which did not consider marriage as a sacrament (Bieler, 1963). The Civil Code appeared to be a solution allowing the same rules to be imposed on all, regardless of confession. However, Catholics were very sensitive to anything that might appear as an attack against the sanctity of the sacrament of marriage (Zogmal, 1998, p. 221). After violent debates a compromise emerged in 1824. From this date on until 1861, two matrimonial regimes coexisted on the cantonal territory. Nevertheless, in all cases, the establishment of a marriage certificate in the secular civil register was compulsory (Oris & Perroux, 2007, pp. 206–207).

While the interconfessional tensions around marriage faded in Geneva, they were revived at the national level, during the first revision of the Swiss federal constitution in 1874, specifically regarding one of its laws of application of direct interest to us, the law on civil registration, marriage and divorce. This legislation was supported by an alliance of radicals and liberals, united against the conservators (mainly Catholic). Its main purpose was to implement Article 47 of the new constitution, which removed the right of the cantons to prohibit the marriage of their indigent citizens and protected interfaith marriages, which many cantons had prohibited, either directly or by raising obstacles. It is this last point in particular that justified the law's imposition of civil marriage and its registration in the civil registry throughout Switzerland (Oris, 2020).

This legislation directly impacted the marriage certificate and its content. The format of the certificates was stable from 1798 (when the French revolutionary armies occupied Geneva) until 1875. Information on 10 persons was provided: for the spouses: last and first names, date of birth and age, matrimonial status and if applicable date of widowhood, birth place, occupation or status, domicile, signature, and mention of illegitimate children if applicable; for the spouses' parents: last and first names, presence or absence at the civil ceremony, occupation, domicile, date of death if applicable, consent, signature; and for four witnesses: occupation, age, kin tie to the spouses, signature. When the 1874 law became effective, the number of witnesses was reduced from four to two, the certificate was simplified and omissions of information, especially on the parents and witnesses, became frequent (Ryczkowska, 2013, pp. 18–19; Schumacher, 2010, p. 336).

A comparison can be done with a third data source which was also affected by the 1874 law: the divorce records. Since 1798, formal divorce certificates can be found in the Geneva civil marriage registers. At that time, the records mentioned four named witnesses, but without any indication of their age, occupation or social status. Overall, divorce records were rare and of bad quality. While the registry clerks were used to correctly record marriages, they rarely had to record a divorce. They seemed to have been unfamiliar with regulations, which resulted in missing data, wrong wording, corrections at the bottom of the certificate. Efforts were made to increase reliability, for example in 1803, when a new form was introduced to avoid the loss of information, but a relapse was quickly observed. In 1813, appeared the "classic" formulation of the certificate of divorce such as found until the beginning of the 1870s. Still, it was not exempt of mistakes either, as of 1817, when in case of mutual consent, no mention was made of the professions and domiciles of the spouses who separated (Oris, 2020).

Figure 1 *Geneva data bases and their relationships*

N.B.: In orange, data which have been linked at the individual level.

As a consequence of the 1874 constitutional reform (see above), the marriage register was composed on the basis of a pre-printed form starting from 1877. There were no more divorce records but the transcription in the margin of a marriage certificate that this union was dissolved by the judgment of the civil court on a given date. That deprived us of any information on the spouse who initiated the procedure, or on the existence of a mutual consent, precisely when a new national law made the divorce by mutual consent possible in the whole of Switzerland (Oris, 2020). Eventually, 518 divorces were digitized for the period 1800–1880, and linked with the marriages.

Another important source for the Geneva project are the residence permits. They were issued by the *Chambre des Etrangers* upon presentation of certificates of origin and good conduct as well as a proof of source of income. These permits had to be renewed every three months and contained a lot of personal information such as name, origin, marital status, names of possible accompanying spouses and children, profession, all places of residence (addresses) in Geneva, and even for two thirds of the persons, their destination after leaving Geneva. Moreover, they concerned both foreigners and confederates from other cantons, given the low degree of integration of the Swiss Confederation before the first federal constitution in 1848. The completeness and reliability of this system was guaranteed by "the will of the Geneva authorities to control the foreign population" (Schumacher, 2010, p. 353), notably by informing newcomers of their duty to declare themselves and by imposing heavy sanctions on offenders as well as on those who might have harbored them (see Schumacher (2010, pp. 350–360) and Remund (2012) for a complete discussion).

While conducting work on immigrations and settlements, Adrien Remund constructed a sample for the timespan between the two censuses of 1837 and 1843. This choice made it possible to take advantage of the censuses to specify the context in which the immigrants evolved. It also allowed studying migration between 1837 and 1843 by linking individuals across the two censuses (Remund, 2010). During the six years of the intercensal interval, 14,489 residence permits were delivered. This means that mobility was intense and that most of the newcomers stayed out rapidly. Among those 14,489, an alphabetical sampling of all persons whose family name begins with the letter "B" was carried out. The sampling rate was about 13%, which represents a sample of 1,903 permits, sufficient to ensure the significance of the analyses performed. In most cases, each new stay generated a new entry in the register. Indeed, although the identification of each duplicate is a very difficult task due to the widespread use of classical first names at that time, as well as approximate spelling and age declaration, an estimate based essentially on name, first name, and year of birth indicates that the 1,903 sampled permits corresponded to about 1,657 distinct individuals, of whom 1,492 made only one trip to Geneva, 117 made two trips, and 48 individuals made up to six consecutive trips during the six-year period (Remund, 2010).

This source suffers from two weaknesses. First, servants were absent of the registers until 1844, although since 1838 male or married servants had to obtain a so-called servant's booklet (Schumacher 2010, p. 248). Second, departure dates are unreliable due to a significant number of migrants who did not bother to retrieve their papers at the end of their stay, and thus do not have an official departure date, nor a destination. The number of permit renewals is therefore a more reliable indicator of time spent in the city, as each renewal corresponds to a three-month extension.

3 RECORD LINKAGES AND STANDARDIZATION

For the sample we opted for an alphabetical approach because it makes much easier the research of additional information on the individuals in various data sources. In that perspective, a crucial point is that in the 19th century Geneva sources, married women did not take the name of their husband but kept their maiden name (the same was true in Belgium; Puschmann, Matsuo, & Matthijs, 2022). We were consequently able to follow women like men, even after their marriage, which avoids a gender bias in the life courses' reconstruction that the use of an alphabetical sample would otherwise imply (Bourdieu, Kesztenbaum, & Postel-Vinay, 2014).

Individuals from one census were linked to the next censuses through a semi-automatic record linkage procedure. Research was done on the four first letters of the surname supplemented with the year of birth (± 2 years). The household of the individual in the original census was displayed on the screen, as was the household of the identified candidate in the next census (t+6, exceptionally t+3). A member of the team then had the opportunity to validate the linkage not only for the individual sought but also for all the individuals in the households displayed. In a city where in- and out-migrations were high, we also searched at t+12, t+18, etc. When in doubt, occupation, location in the city and religion were used to assess the plausibility of the linkage. Table 1 displays the results, showing only the direct links. For example, among the 2,909 people belonging to the B sample in 1816, 22 were not found in the subsequent censuses but could be linked in 1843, 27 years later. The loss of a register in 1831 impacted both the sample size and the linkage rate, which otherwise varied between 52 and 60%.

Our simple approach using the household context revealed to be very efficient to overcome many variations in the individual data. Out of the total of 7,624 linkages, differences in surname spelling were observed in no less than 35.5% of the cases. Differences in first names were even more frequent, either because of the use of abbreviations, or because sometimes only the first one was mentioned instead of two or three in another census.

Table 1 *Size of the B sample and number of links with subsequent appearances across the six censuses from 1816 to 1843*

Censuses	B sample	Linkages with individual notices in				
		1822	1828	1831	1837	1843
1816	2,909	1,469	149	85	34	22
1822	2,957		1,411	122	55	5
1828	3,209			788	93	209
1831	2,936				1,377	88
1837	3,315					1,717
1843	3,650					
Total	18,976	1,469	1,560	995	1,559	2,041

The linkages' robustness was assessed through fully automated tests of likelihood and tests of consistency. The former concerned the plausibility of the age of the spouses, the age gap between them, the age of the parents at the birth of their children, the age differences between siblings, and the 'elite' socioeconomic status when the individual was younger than 25. Such tests are helpful to identify false positives, i.e., plausible links that do not survive to a critical check. Consistency checks reveal errors in the data and/or in the linkages by focusing on changes across time in the same variables. Typically, sex, place of birth and place of origin must remain constant, and age must change coherently. Biographical transitions have also to be consistent (a married person cannot become single), as well as changes in cohabitation. Around 220 linkages were found to be erroneous. Most of these were people with very common first and last names, as well as a few twins.

The linkage of the individual appearances across the censuses was part of a collective project. Other sources were linked by individual researchers. For his master, then PhD thesis, Reto Schumacher (2002; 2010) built his own database, realizing a tedious work of data linkages to produce a massive family reconstitution in a 19th-century urban context. He first used the sample of Grazyna Ryczkowska (see above) to make a subsample of 2,573 marriages celebrated in Geneva between 1800 and 1880 of all couples among which the husband's name began with a B. The sample included about 12.5% of the marriage records (12.3 for the period 1800–1846, 12.6 for the years 1847–1880). Using the registers of residence permits (previously presented), he further collected information on 1,666 immigrant married couples where the husbands' name started with a B and who settled in the city between 1844 and 1880. Relying on the birth and death certificates from civil registration he then went on to reconstruct the couples' reproductive histories in Geneva (Schumacher, 2010, pp. 331–341). Adapting Louis Henry's method of family reconstitution, Schumacher also used the registers of residence permits, population censuses and address books, to precisely establish the dates of beginning and end of observation, and the types of censoring to the left or to the right. This proved impossible for 647 families who were excluded. He eventually reconstructed the trajectories of 3,592 families, offering the opportunity to compare the city natives with immigrants (see Schumacher, 2010, chapter 9 for a full discussion).

Grazyna Ryczkowska also performed linkages across marriages. Among the 8,507 marriages involving a "B" that took place in the canton of Geneva between 1800 and 1880, 6,346 celebrated from 1830 onwards were selected and the parents of the "B" spouse were researched in the whole database, thus from 1800. While approximate spelling was a frequent issue in the censuses and in the registration of residence permits, marriage certificates were almost perfect from this point of view. This was probably due to the fact that the marriage certificate had a legal value and also that the engaged couple had to provide an official copy of their birth or baptism certificate which also included the name of the parents. Within this sample some 1,372 links (21.6%) were easily established across generations following this methodology (Ryczkowska, 2013, p. 182).

In all sources, all the information was fully digitized, so that each researcher could use his/her own method of codification, typically of locations and occupations. Across the various data sources, locations have been coded through a classification centered on Geneva. In addition to the city itself, the classification distinguished the peri urban municipalities as well as the catholic and the protestant countryside. Beyond the borders of the Geneva canton, the canton of Vaud was kept distinct because of its important ties with Geneva, then the rest of French-speaking Switzerland, and eventually the rest of Switzerland (German- and Italian-speaking regions). A region still known today as "neighboring France" was also kept distinct from the rest of France. Italy and Germany, although not unified until 1870–1871, and a residual group of "other countries" complete this classification. Only 1.7% of the locations stayed indeterminate (Ryczkowska, 2013, pp. 19–22). Reto Schumacher, however, used a different classification because he wanted to compare the fertility of migrants in Geneva with the fertility in their region of origin in a multilevel analysis. He thus adapted his geographical typology to the availability of the Ig index of legitimate fertility (Coale & Watkins, 1986; Schumacher, 2010, pp. 473–477).

Regarding occupations, the debates initially opposed Marxist and non-Marxist classification systems, and more recently focused on comparability over time and countries versus specificity. This discussion will probably never end, and ultimately it comes down to the researcher(s) to be transparent about their objectives and the tools they use or suggest to reach their aims. Most of us used a classification which distinguished the unskilled workers, the blue collars, the "Fabrique" (watch-and-clock-makers), the white collars, the petty bourgeoisie and the elite. It was an adaptation of SOCPO (Van de Putte & Miles, 2005), the peculiarity being that the "Fabrique" was isolated, considering its importance in Geneva. Following a suggestion of Guy Brunet, it was tested whether all siblings belonged to the same

or the adjacent classes and the results were satisfactory, suggesting that this classification is robust (Ryczkowska, 2013, pp. 22–27). Schumacher (2010, pp. 444–448) suggested a more sophisticated approach, coding first the occupations with HISCO (van Leeuwen, Maas, & Miles, 2002), and on this basis with HISCLASS creating 11 categories (van Leeuwen & Maas, 2011), before doing an analysis of homogamy (on the class of the father of the groom and the class of the father of the bride) to measure social interactions. This approach led to a reduction of the eleven HISCLASS categories to four classes: elite, small shopkeepers and white collars, skilled workers and craftsmen, unskilled workers.

Census data have also been coded according to the Hammel and Laslett (1974) typology of household structures. Our team further developed a typology of solitude, from the most obvious situations (living alone) to more ambiguous (spinsters and bachelors, widows and widowers, Lodgers, servants) (see Oris, Ritschard, and Ryczkowska (2006a) for a full discussion). All the classifications that have been developed are available for researchers who would like to use or modify them.

4 IMPACT

Compared to the massive demographic reconstruction carried out in Quebec, Sweden, the Netherlands or China (resp. Vézina & Bournival, 2020; Dribe & Quaranta, 2020; Edvinsson & Engberg, 2020; Mandemakers & Kok, 2020; Campbell & Lee, 2020), the Geneva database is far more modest in size and coverage, and could not expect a similar impact. However, the promises have been kept and the future remains open.

4.1 DEMOGRAPHIC AND FAMILY SYSTEMS

Analyses of the collected material revealed an original urban demographic regime. The signs of modernity were clear, with low marital fertility and infant mortality prevalent since the early 19th century. Birth control was obvious, as shown by a TFR of 2.32 children for couples married between 1800 and 1850, and a risk of dying before the first birthday between 100 and 130‰ in the first half of the 19th century (Schumacher, 2010). This makes the whole city of Geneva a pioneer population in the demographic transition, both from a Swiss and European perspective. However, Geneva combined those modern traits with traditional ones. Indeed, the average age at first marriage was 28 for women and final celibacy was above 20% (Ryczkowska, 2013). From this point of view, although the figures recorded in Geneva were particularly high, the city can be seen as just an example of the North-Western Europe demographic system where access to marriage was restrained and where the nuclear family form was dominant (Hajnal, 1982; Laslett, 1983). Indeed, the large majority of this urban population (61% of individuals) lived in nuclear households. Extended and multiple households were scarce (respectively 6.3% and 8.4%).

4.2 VULNERABLE POPULATIONS, VULNERABLE WOMEN

This classical typology of household forms hides the many residents who did not belong to a nuclear family: the cohabitants and the lodgers, as well as the servants and the seasonal or temporary workers. They have been studied in a line of research that focused on vulnerable individuals in an urban environment. Although various exposures to loneliness were highly prevalent in Geneva, especially among women, with in addition a frequent accumulation of disadvantages (Oris et al., 2006a), solitude was rare: only 6.4% of the urban residents lived alone. Older adults offer a good illustration of this paradox. Following the nuclear hardship hypothesis, they were expected to end their life in an empty nest, abandoned by their children. Indeed, young Genevans who grew up in a local urban family moved out late, and did so directly from the parental home to their own neo-local household (Oris, Ritschard & Ryczkowska, 2005). Older women were much more at risk to end their life alone because they were more often single (final celibacy reached 20% among women against 10% among men) and widows (accounting for 45% of women against 25% of men in the age group 55+). Women, and especially those living without a husband, were moreover concentrated in low-income occupations. However, one of the sources of vulnerability was also a solution: the daughters who "sacrificed" themselves by staying single to take care of their old parents. Additionally, married children, more often sons, hosted older parents in their household, but it seems that this solution only applied when parental health was severely impaired. Alternative solutions, especially for old women,

were cohabitation (usually with siblings) and becoming lodgers. Despite none of these situations being ideal, they allowed more than 85% of elderlies to avoid solitude. Moreover, the study of the turnover based on the linkages of the censuses' individual records has shown that spinsters and widows aged 45 and older were among those who stayed the most in Geneva, very probably because they could benefit from the urban welfare institutions (Ryczkowska & Perroux, 2006).

Those vulnerable but stable adult women contrasted with a very mobile group of teenagers and young adults from rural families who came to Geneva as servants and labourers, providing the city an important number of single migrants aged 15 to 35. As in many preindustrial towns, women engaging in domestic service and various personal services (cleaning, ironing, etc.) made up the highest share of these young migrants, generating an unbalanced marriage market (70 men for 100 women at 20–24) (Oris et al., 2006a). This imbalance contributed to maintain a late age at first marriage and a high prevalence of final celibacy among women (Ryczkowska, 2013). For many historians, immigrants and especially maids appeared to be highly vulnerable. In 19th century Geneva like elsewhere, contemporary observers saw them as the source of all evils, a real threat to morality. Confirming both old stereotypes and previous studies, Schumacher, Ryczkowska, and Perroux (2007) have shown that both the marriage market and female poverty played an important role in forcing young women to engage in premarital sexual relations. However, servants did not experience a higher risk of out-of-wedlock births compared to other unskilled women, no more than immigrant women compared to native-born women. Contrasting with the fears of the religious and public authorities, during the first half of the 19th century, the evolution of child abandonment, out-of-wedlock births, and premarital conception, together show a reinforcement of social control.

4.3 MANY CAME, FEW STAYED

These results show that we should refrain from drawing too rapid conclusions about the vulnerability of young migrants, that miserabilism is not the right approach, even in a society where poverty and inequalities were widespread. A city like Geneva was not a factory of vulnerabilities but a crossroad, attracting but also rejecting a lot of people. At the bottom of social structures, the migratory turbulence was extreme. While unskilled workers made up 31% of the population declaring an occupation in a given census, barely one out of five was still holding the same status 6 years later (Oris & Ritschard, 2007). Staying and settling durably was a very selective process. Most of the immigrants came from neighboring France and French-speaking Switzerland, but also from German-speaking Switzerland and Germany, and from northern Italy (Remund, 2009; Ryczkowska, 2013; Schumacher, 2010). Adrien Remund (2010; 2012) has shown that a third of the migrants left after three months, half of them after one year. Later on, departures remained high until reaching approximately 10% of the initial cohort of immigrants. This small minority of migrants, about 250 per year, represented the share that eventually settled in Geneva (Remund, 2013; 2014).

A figure of 250 migrants seems modest, but it was responsible for 94% of Geneva's demographic growth, since birth control and restrained access to marriage resulted in a very low natural balance of births and deaths. This significantly impacted the social structure of Geneva. In the early modern period, many European cities (Le Roy Ladurie, 1998; Lynch, 2003) were under the control of a rooted segment of the population that preserved its political and socioeconomic domination by defining and attributing various statutes of residence: inhabitants, natives, and bourgeois. In Geneva, tensions between those groups degenerated into violent episodes during the 18th century. The occupation of the city by the French revolutionary armies, and later the adhesion to the Swiss Confederation, resulted in a new form of citizenship and constrained the openness to migrants. As the population of their city changed due to the migratory flows that accelerated from 1798 onwards, the natives of Geneva tended to define themselves as the "old Genevans" (Herrmann, 2003). In a city where "Genevan" remained the nationality and the primary collective identity (Remund, 2009), the authorities maintained a strict control of the "foreigners" (Remund, 2010).

4.4 THE CATHOLIC QUESTION

However, forced to respect international treaties signed by the Swiss Confederation, and because from 1816 on, Geneva had to mutate to a religiously and geographically mixed canton, the authorities could not reject the Catholics. The arrival of these Catholic migrants among the newcomers was quite a shock since during centuries, the Geneva city-state constructed its identity as the "Protestant Rome". The settlement of Catholics was prohibited until 1798. When introducing the marriage certificates, we

have mentioned the resistance of the Protestant elites towards this change. Other qualitative evidence suggests vivid tensions (Oris & Perroux, 2007).

The quantitative analysis of the collective biographies reconstructed through the linkages of the individual records in the population censuses, suggests however a different story, more peaceful, less confrontational. As early as 1816, Catholics made up 11% of the city population. This proportion grew to 28% in 1843 and 46% in 1900. Initially, most of them were young single adults engaged in labor migration, with a high turnover rate. In the 1820s, conversions to protestantism and education of the offsprings of mixed marriages in the protestant faith threatened their survival as a minority. However, the age structures progressively changed, families settled durably, the first catholic children born in Geneva since 1536 grew up, aged and eventually died in the city. In the 1830s and 1840s, the size of the Catholic population was already sufficient to offer an internal marriage market (Remund, 2009; Ryczkowska, 2013). That does not mean, however, that the Catholics in Geneva developed as a closed community surrounded by an aggressive majority. On the contrary, in the first half of the 19th century mixed households were continuously more numerous than the catholic homogeneous ones (Oris & Perroux, 2007). Additionally, implicative statistics was used to identify possible social and economic discriminations. Results showed that Catholics were not concentrated in a given social class or in specific branches of activities (Oris, Ritschard, & Perroux, 2010; 2013). Similarly, spatial segregation was inexistent: no catholic neighborhood could be identified, not even a catholic street (Remund, 2010; 2012). Far from the history of Chinatowns, little Italies, and other ghettos made popular by the School of Chicago (Laurie & Khan, 2017), Catholics in Geneva "lived hidden in plain sight to live in peace". This strategy is widespread among minorities despite the scientific literature often focusing on the experiences of discrimination.

The Catholic and non-Catholic migrants who were staying in the city, formed an intermediate group between the rooted Genevans and the more mobile part of the immigration. Through a cumulative effect, non-natives slowly but surely took an increasingly important place in the city's population, and each new wave of immigrants who passed through "the urban labor market without any intention of lasting integration, at least at first", now found their bearings there (Oris & Perroux, 2007, p. 226). The city of Geneva illustrates in this sense the well-known process of migration chains (Remund, 2012). As a result of a narrow selection process, these new Genevans knew how to keep a low profile in a hostile city, preferring to be forgotten until they were eventually considered legitimate Genevans.

4.5 THE PROTESTANT STABILITY POLES

Doing so, Catholics and other newcomers grew in number without threatening the domination of the Protestants natives who were occupied with their own divisions. Statistics indeed show that Protestants were overrepresented in the bourgeoisie and in the "Fabrique" (Oris, Ritschard, & Ryczkowska, 2006b). As mentioned in the introduction, those two groups fought for power until the 1840s. The watch-and-clock makers were by far the most numerous. More than a third of the grooms marrying in Geneva between 1822 and 1845 were active in this urban proto-industry (Ryczkowska, 2003). Those workers with highly specialized skills formed an aristocracy of blue collars who distinguished themselves from the other skilled manual workers and shopkeepers, all actors of an urban "molecular capitalism". Deeply rooted in Geneva (Ritschard, Studer, & Oris, 2009), the craftsmen engaged in the "Fabrique" transmitted their status across the generations. Building on the linkages of the marriage certificates and using mobility trees (Ritschard, Studer, Müller, & Gabadinho, 2007), Grazyna Ryczkowska (2013, Chapter 5) has shown that having a watch-and-clock maker grandfather strongly predicted the grandson's belonging to the watch-and-clock-maker industry. Children who grew up in those families were 30% less at risk of leaving Geneva than the offspring of other blue collars (Oris et al., 2005). Concentrated in a neighborhood, literate like all the Protestants, politicized, they made most of the revolutionaries who in 1846 gave the power to the Radical party. The bourgeoisie was much less numerous but also mainly made of families who preserved their status across centuries and were allied through repeated homogenous marriages. Because of their engagement in the "Protestant diaspora" active in trade and finance, they seemed to be more mobile and to have an international marriage market. In reality however, they stayed an essentially closed group (Perroux, 2006). Both the watch-and-clock makers and the bourgeoisie constituted what Emmanuel Le Roy Ladurie (1998, p. 301) called the "stability poles" of this urban society.

4.6 SOCIAL STRUCTURES AND DEMOGRAPHIC BEHAVIORS

This stability of the urban social structures facilitates research on differential demographic behaviours in a long-term perspective. Alfred Perrenoud (1975) has demonstrated the existence of impressive differences in mortality among the social groups in 17th century Geneva. Variations from the simple to the double between the elite and the popular classes were mainly due to infant and child mortality, and differential exposure to smallpox. Such inequalities in children survival however already decreased in the 18th century, and in the 19th century popular classes and skilled workers were approximately at the same level. Elites were still favored, but much less than in the previous centuries. Geneva is a unique place to reconstruct long-term trends, which support both the constancy hypothesis (constant elite advantage or fundamental cause theory) and the convergence hypothesis on social inequality in death (Schumacher & Oris, 2011).

During the 19th century, the probability of surviving to age 5 reached 87% among the children from the upper classes (against 77% for the children of manual workers) (Schumacher, 2010, Chapter 11). In Geneva like elsewhere, infant deaths were clustered in specific families and elites were overrepresented in the low infant mortality group (Schumacher, 2016, p. 105). Upper classes could consequently drastically control their births through both spacing and stopping practices, without threatening their reproduction. If the elites' marital fertility was particularly low compared to the other socioeconomic groups, it was however in a context where birth control was generalized. During most of the 19th century, Geneva showed a transitional level (approximately 40% of the Hutterite fertility) (Schumacher, 2013, p. 157).

A plausible explanation of this stable and low fertility resides in the transformation of the population structures, more precisely in the progressive accumulation of immigrants among the city inhabitants. Newcomers played an important role in the heterogeneity of demographic behaviors. Reto Schumacher (2010; 2013) has shown that being born in a region where the index of legitimate fertility (Ig) was above .600 resulted in a 50% higher fertility for the migrant couples who arrived in Geneva already married. This strong impact of their initial socialization was attenuated among those who immigrated relatively young, before being aged 30, suggesting that living in Geneva and observing the native families could result in an evolution of the models and values. This was visible among the migrants who married in Geneva. Even those coming from high-fertility regions adapted to a large extent to the low fertility of the Geneva natives, only limited socialization effects staying apparent (Schumacher, Matthijs, & Moreels, 2013).

5 SUMMARY

This project started in 2003 with the population censuses. In historical demography, studies based on this type of data source became mainstream in the 1970s and early 1980s, thanks to the implementation of mainframe computers in social sciences research. This success was made stronger by the fact that this intellectual period was dominated by the structuralist thought. A generation of researchers faced the challenge of dealing with massive data on 19th century growing urban populations (see, for example, Desama, 1985; Guillaume, 1972; Hershberg, 1976). Later, new approaches emerged focusing more on the dynamics of changes in urban social and demographic structures and behaviors (Bourdelaïs & Demonet, 1995; Reher, 1990). Those researchers changed scale to analyze urban populations at the level of households (Janssens, 2002; Laflamme, 2007), families (Eggerickx, 2004; Faron, 1997; Pétilion, 2006) or individuals (Alter, 1988; Kertzer & Hogan, 1989).

Adding more sources and systematically linking individual data was crucial to reconstruct "collective biographies" and trajectories across the city (Pinol, 1999). Especially in countries without population registers (Breschi, Fornasin, & Manfredini, 2020; Sommerseth & Thorvaldsen, 2022), researchers face the challenge to construct longitudinal data in turbulent contexts, but can relatively often rely on a wealth of documents produced by various urban administrations (see Paping & Sevdalakis, 2022; Puschmann, Matsuo, & Matthijs, 2022; or the recent Charleville database described in Alexandre, Dupuy, & Gourdon, 2022). Drawing also from the diffusion of analytical methods (especially event-history-analysis) that made possible the analysis of this new generation of databases (Alter, 1998), our Geneva project tried to take the best from all those experiences. Also inspired by the

pioneering "Philadelphia Social History Project" (Hershberg, 1976), we paid a specific attention to the embeddedness of historical demography and social history, which is far from being as obvious as it might seem (see Cahen & Kesztenbaum, 2019). Less popular methods in our fields, such as implicative statistics (Gras, Suzuki, Guillet, & Spagnolo, 2008), inductive trees or data mining tools (Ritschard & Oris, 2005), were useful in that perspective. The confrontation of qualitative data sources and the social representations they allowed to depict with the results from various quantitative analyses also helped fulfilling this goal.

Looking back at the development of our project(s), we have to acknowledge a certain dose of adaptation to circumstances. Originally based on the population censuses of the first half of the 19th century, thanks to the support of the Swiss National Science Foundation¹, our aim was to systematically add the births and deaths to the B-sample, as well as the residence permits for foreigners and passports for Genevans, so to reconstruct quasi population registers. But we were short of funds and this work has not been achieved, except for the illegitimate births (see Schumacher et al., 2007).

The decision to opt for an alphabetical sample was however decisive. Later on, research assistants of the project and other students, for their PhD and master theses, were able to build on those foundations and support each other while growing also as individual projects. Adding the marriages was a tremendous gain. Still today and in the future, new research questions can interrogate the data, and new data can be added at any time. Censuses, marriages, divorces, and residence permits' databases are available to the scientific community unconditionally, including the data that have been coded and the classifications used. Dbase and Access have been used, but transfer in another format can be easily managed.

Beyond the scientific community, it is worth mentioning that the diverse databases that were described above were also heavily exploited in the creation of the *poliscope*, the outreach project of the Faculty of Social Sciences of the University of Geneva.² Building on the academic works mentioned above, as well as the digitalization of the Relief Magnin, a three-dimensional representation of the city just before the demolition of the fortifications was created. The *poliscope* consists of a series of lectures and round tables mainly geared towards high-school students, which includes among others game-like software. Visitors are allowed to navigate through the Geneva of the 1850s and meet fictive characters directly inspired from real groups of migrants from that period. In Geneva nowadays more than 40% of the population are migrants from approximately 190 countries, and they are particularly numerous among the teenagers. In this context, this initiative proved highly successful among young crowds, notably to initiate reflections on topics such as migrant integration, equal rights and gender and religious discriminations.

REFERENCES

- Alexandre, C., Dupuy, J., & Gourdon, V. (2022). Nouveaux regards sur Charleville [New perspectives on Charleville], Charleville-Mézières, Société d'histoire des Ardennes. In *Cahier d'études ardennaises* (Vol. 26). Retrieved from <https://hal.science/hal-03905374>
- Alter, G. (1988). *Family and the female life course: The women of Verviers, Belgium, 1849–1880*. Madison: University of Wisconsin Press.
- Alter, G. (1998). L'event history analysis en démographie historique. Difficultés et perspectives [Event history analysis in historical demography. Problems and prospects]. *Annales de Démographie Historique*, 2, 25–35. doi: [10.3406/adh.1999.1934](https://doi.org/10.3406/adh.1999.1934)
- Bardet, J.-P. (1983). *Rouen aux XVIIe et XVIIIe siècles: Les mutations d'un espace social* [Rouen in the 17th and 18th centuries: The mutations of a social space]. Paris: Sedes-CDU.
- Bardet, J.-P., & Dupâquier J. (Eds.). (1998). *Histoire des populations de l'Europe* [History of the European populations] (Vols. 1-3). Paris: Fayard.
- Bengtsson, T., & van Poppel, F. (2011). Socioeconomic inequalities in death from past to present: An introduction. *Explorations in Economic History*, 48(3), 343–356. doi: [10.1016/j.eeh.2011.05.004](https://doi.org/10.1016/j.eeh.2011.05.004)

1 Swiss National Science Foundation grants 1114-68113 and 100012-105478. This support is gratefully acknowledged.

2 <https://poliscope.ch/projets/installation-multimedia/destins-croises-des-migrants-dhier/>

- Bieler, A. (1963). *L'homme et la femme dans la morale calviniste: La doctrine réformée sur l'amour, le mariage, le célibat, le divorce, l'adultère et la prostitution, considérée dans son cadre historique* [Man and woman in Calvinist morality: The Reformed doctrine on love, marriage, celibacy, divorce, adultery and prostitution, considered in its historical context]. Genève: Labor et Fides.
- Bourdelaïs, P., & Demonet, M. (1995). L'industrialisation: L'exemple du Creusot essai d'histoire des itinéraires individuels (1836–1881) [Industrialisation: The Creusot case. Essay on the history of individual itineraries (1836–1881)]. *Les Cahiers du Centre de Recherches Historiques*, 14–15, 1–5. doi: [10.4000/ccrh.2663](https://doi.org/10.4000/ccrh.2663)
- Bourdieu, J., Kesztenbaum, L., & Postel-Vinay, G. (2014). *L'enquête TRA, histoire d'un outil, outil pour l'histoire: Tome I. 1793–1902* [The TRA survey, the history of a tool, a tool for history]. Paris: INED.
- Breschi, M., Fornasin, A., & Manfredini, M. (2020). The richness of Italian historical demography. *Historical Life Course Studies*, 9, 228–240. doi: [10.51964/hlcs9304](https://doi.org/10.51964/hlcs9304)
- Cahen, F., & Kesztenbaum, L. (2019). Introduction: Pour un dialogue entre démographie historique et histoire sociale [Introduction: Toward a dialogue between historical demography and social history]. *Annales de démographie historique*, 2, 7–20. doi: [10.3917/adh.138.0007](https://doi.org/10.3917/adh.138.0007)
- Campbell, C., & Lee, J. (2020). Historical Chinese microdata. 40 years of dataset construction by the Lee-Campbell research group. *Historical Life Course Studies*, 9, 130–157. doi: [10.51964/hlcs9303](https://doi.org/10.51964/hlcs9303)
- Coale, A. J., & Watkins, S. C. (1986). *The decline of fertility in Europe*. Princeton: Princeton University Press.
- Desama, C. (1985). *Population et révolution industrielle: Évolution des structures démographiques à Verviers dans la première moitié du 19e siècle* [Population and industrial revolution. Evolution of the demographic structures in Verviers during the first half of the 19th century]. Genève: Librairie Droz.
- Dribe, M., & Quaranta, L. (2020). The Scanian Economic-Demographic Database (SEDD). *Historical Life Course Studies*, 9, 158–172. doi: [10.51964/hlcs9302](https://doi.org/10.51964/hlcs9302)
- Dupâquier, J. (1984). *Pour la démographie historique* [For historical demography]. Paris: Presses Universitaires de France.
- Edvinsson, S., & Engberg, E. (2020). A database for the future. Major contributions from 47 years of database development and research at the Demographic Data Base. *Historical Life Course Studies*, 9, 173–196. doi: [10.51964/hlcs9305](https://doi.org/10.51964/hlcs9305)
- Eggerickx, T. (2004). *La dynamique démographique et la transition de la fécondité dans le bassin industriel de la région de Charleroi, de 1831 à 1910* [Demographic dynamics and the fertility transition in the industrial basin of Charleroi region, from 1831 to 1910]. Bruxelles: Académie royale de Belgique. Retrieved from <http://hdl.handle.net/2078.1/155147>
- Faron, O. (1997). *La ville des destins croisés. Recherches sur la société milanaise du XIXe siècle (1811–1860)* [The city of crossed destinies. Research into 19th-century Milanese society (1811–1860)]. Rome: École Française de Rome.
- Gras, R., Suzuki, E., Guillet, F., & Spagnolo, F. (Eds.). (2008). *Statistical implicative analysis. Theory and applications*. Berlin: Springer-Verlag.
- Guillaume, P. (1972). *La population de Bordeaux au 19e siècle* [The population of Bordeaux in the 19th century]. Paris: Armand Colin.
- Hajnal, J. (1982). Two kinds of preindustrial household formation system. *Population and Development Review*, 8(3), 449–494. doi: [10.2307/1972376](https://doi.org/10.2307/1972376)
- Hammel, E. A., & Laslett, P. (1974). Comparing household structure over time and between cultures. *Comparative Studies in Society and History*, 16(1), 73–109. doi: [10.1017/S0010417500007362](https://doi.org/10.1017/S0010417500007362)
- Hatt-Diener, M.-N. (2004). *Strasbourg et Strasbourgeois à la croisée des chemins. Mobilités urbaines, 1810–1840* [Strasbourg and the people of Strasbourg at a crossroads. Urban mobility, 1810–1840]. Strasbourg: Presses universitaires de Strasbourg.
- Henry, L. (1956). *Anciennes familles genevoises. Étude démographique: XVIe–XXe siècles* [Old Genevan families. Demographic study: 16th–20th centuries]. Paris: Presses universitaires de France.
- Henry, L., & Gauthier, E. (1958). *La population de Crulai, paroisse normande* [The population of Crulai. A parish in Normandie]. Paris: Presses universitaires de France.
- Herrmann, I. (2003). *Genève entre République et Canton: Les vicissitudes d'une intégration nationale (1814–1846)* [Geneva between Republic and Canton: The vicissitudes of national integration (1814–1846)]. Genève & Québec: Editions Passé présent & Presses de l'Université Laval.

- Hershberg, T. (1976). The Philadelphia social history project: An introduction. *Historical Methods Newsletter*, 9(2–3), 43–58. doi: [10.1080/00182494.1976.10112634](https://doi.org/10.1080/00182494.1976.10112634)
- Janssens, A. (2002). *Family and social change. The household as a process in an industrializing community*. Cambridge: Cambridge University Press.
- Kertzer, D. I., & Hogan, D. P. (1989). *Family, political economy, and demographic change: The transformation of life in Casalecchio, Italy, 1861–1921*. Madison: University of Wisconsin Press.
- Laflamme, V. (2007). *Vivre en ville et prendre pension à Québec aux XIXe et XXe siècles* [Living in town and boarding in Quebec in the 19th and 20th centuries]. Paris: L'Harmattan.
- Laslett, P. (1983). Family and household as work group and kin group: Areas of traditional Europe compared. In R. Wall, J. Robin & P. Laslett (Eds.), *Family forms in historic Europe* (pp. 513–563). Cambridge: Cambridge University Press.
- Laurie, T., & Khan, R. (2017). The concept of minority for the study of culture. *Continuum*, 31(1), 1–12. doi: [10.1080/10304312.2016.1264110](https://doi.org/10.1080/10304312.2016.1264110)
- Le Roy Ladurie, E. (1998). *La ville des temps modernes de la Renaissance aux Révolutions* [The early modern time city, from the Renaissance to the Revolutions]. Paris: Seuil.
- Lorenceau, R. (2001). *Bâle de 1860 à 1920: Croissance et mobilités urbaines* [Basel between 1860 and 1920: Urban growth and urban mobility] (PhD thesis). University of Tours. Retrieved from https://ipna.duw.unibas.ch/fileadmin/user_upload/ipna_duw/PDF_s/BBS_PDF/Lorenceau_Diss_2001.pdf
- Lynch, K. A. (2003). *Individuals, families and communities in Europe, 1200–1800. The urban foundations of Western society*. Cambridge: Cambridge University Press.
- Mandemakers, K., & Kok, J. (2020). Dutch liives. The Historical Sample of the Netherlands (1987–): Development and research. *Historical Life Course Studies*, 9, 69–113. doi: [10.51964/hlcs9298](https://doi.org/10.51964/hlcs9298)
- Oris, M. (2020). Divorcer à Genève au XIXe siècle. Analyse sociodémographique d'une rupture contractuelle [Divorce in 19th-century Geneva. A sociodemographic analysis of a broken contract]. *Annales de Démographie Historique*, 2, 107–128. doi: [10.3917/adh.140.0107](https://doi.org/10.3917/adh.140.0107)
- Oris, M., & Perroux, O. (2007). La minorité catholique dans la Rome protestante. Contribution à l'histoire démographique de Genève dans la première moitié du XIXe siècle [The Catholic minority in Protestant Rome. Contribution to the demographic history of Geneva in the first half of the 19th century]. In J.-P. Poussou & I. Robin-Romero (Eds.), *Histoire des familles, de la démographie et des comportements, en hommage à Jean-Pierre Bardet* [History of families, demography and behaviors. A tribute to Jean-Pierre Bardet] (pp. 201–226). Paris: Presses de l'Université Paris-Sorbonne.
- Oris, M., Ritschard, G. (2007). Dynamique socioprofessionnelle dans la Genève du 19e siècle, enseignements d'une analyse de statistique implicative [Socio-professional dynamics in 19th-century Geneva, lessons from an implicative statistics analysis]. In R. Gras, P. Orus, B. Pinaud & P. Gregori (Eds.), *Nouveaux apports théoriques à l'analyse statistique implicative et applications. 4èmes Rencontres internationales d'analyse statistique implicative* [New theoretical contributions to implicative statistical analysis and applications. 4th International conference on implicative statistical analysis] (pp. 287–300). Castellaon: Unisitat Jaume I.
- Oris, M., Ritschard, G., & Perroux, O. (2010). Growing religious pluralism in early nineteenth-century Geneva: New methods for revealing hidden structures and dynamics from censuses. *Popolazione e Storia*, 11(2), 43–58. doi: [10.4424/ps2010-10](https://doi.org/10.4424/ps2010-10)
- Oris, M., Ritschard, G., & Perroux, O. (2013). Le pluralisme religieux croissant de Genève dans la première moitié du XIXe siècle. Une exploration des dynamiques sous-jacentes [Geneva's growing religious pluralism in the first half of the 19th century. An exploration of the underlying dynamics]. In F. Amsler & S. Scholl (Eds.), *L'apprentissage du pluralisme religieux. Le cas genevois au XIXe siècle* [Learning about religious pluralism. The case of 19th-century Geneva] (pp. 41–61). Genève: Labor et Fides.
- Oris, M., Ritschard, G., & Ryczkowska, G. (2005). Siblings in a (neo-)Malthusian town. From cross-sectional to longitudinal perspectives. *Historical Social Research*, 30(3), 171–194. doi: [10.12759/hsr.30.2005.3.171-194](https://doi.org/10.12759/hsr.30.2005.3.171-194)
- Oris, M., Ritschard, G., & Ryczkowska, G. (2006a). Les solitudes urbaines. Structures et parcours dans la Genève des années 1816–1843 [Urban solitudes. Structures and trajectories in Geneva 1816–1843]. *Annales de Démographie historique*, 1, 59–87. doi: [10.3917/adh.111.0059](https://doi.org/10.3917/adh.111.0059)
- Oris, M., Ritschard, G., & Ryczkowska, G. (2006b). Recrutement et renouvellement des groupes socioprofessionnels à Genève, 1816–1843 [Recruitment and renewal of socio-professional groups in Geneva, 1816–1843]. *14e Colloque de l'Association Internationale des Démographes de Langue Française AIDELF*, 791–805. Paris: AIDELF.

- Paping, R., & Sevdalakis, D. (2022). The Groningen Integral History Cohort Database. Development, design and output. *Historical Life Course Studies*, 12, 78–98. doi: [10.51964/hlcs12033](https://doi.org/10.51964/hlcs12033)
- Perrenoud, A. (1975). L'inégalité sociale devant la mort à Genève au XVIIe siècle [Social inequality in the face of death in 17th-century Geneva]. *Population*, 30(1), 221–243. doi: [10.2307/1530652](https://doi.org/10.2307/1530652)
- Perrenoud, A. (1979). *La population de Genève du seizième au début du dix-neuvième siècle* [The population of Geneva from the 16th to the early 19th century]. Genève: Librairie Droz.
- Perrenoud, A. (1988). Espacement et arrêt dans le contrôle des naissances [Spacing and stopping in the birth control]. *Annales de démographie historique*, 59–78. Retrieved from <https://www.jstor.org/stable/44384928>
- Perroux, O. (2006). *Tradition, vocation et progrès. Les élites bourgeoises de Genève (1814–1914)* [Tradition, vocation and progress. The bourgeois elites of Geneva (1814–1914)]. Genève: Slatkine.
- Pétillon, C. (2006). *La Population de Roubaix: Industrialisation, démographie et société 1750–1880* [The population of Roubaix: Industrialisation, demography and society 1750–1880]. Lille: Presses Universitaires du Septentrion. doi: [10.4000/books.septentrion.55914](https://doi.org/10.4000/books.septentrion.55914)
- Pinol, J.-L. (1999). Faire son chemin dans la ville. La mobilité intra-urbaine [Making your way through the city. Intra-urban mobility]. *Annales de Démographie historique*, 1.
- Prost, A. (2022). *Orléans en 1911. Sociologie d'une ville* [Orléans in 1911. Sociology of a town]. Paris: CNRS.
- Puschmann, P., Matsuo, H., & Matthijs, K. (2022). Historical life courses and family reconstitutions. The scientific impact of the Antwerp COR*-database. *Historical Life Course Studies*, 12, 260–278. doi: [10.51964/hlcs12914](https://doi.org/10.51964/hlcs12914)
- Ramiro Fariñas, D., & Oris, M. (Eds.). (2016). *New approaches to death in cities during the health transition*. Springer International Publishing.
- Reher, D. S. (1990). *Town and country in pre-industrial Spain: Cuenca, 1540–1870*. Cambridge: Cambridge University Press.
- Remund, A. (2009). *Les chemins de la migration. Une analyse de la mobilité étrangère à Genève (1837–1843)* [The migration roads. An analysis of foreign mobility in Geneva (1837–1843)] (Master thesis). University of Geneva. Retrieved from <https://poliscope.ch/files/2017/01/remund-2009.pdf>
- Remund, A. (2010). *Socioeconomic mobility of immigrants in 19th-century Geneva. Confronting cross-sectional and longitudinal approaches* (Master thesis). Lund university. Retrieved from https://www.unige.ch/sciences-societe/ideso/files/4414/3636/9586/EDSD_thesis.pdf
- Remund, A. (2012). Rester ou repartir ? Une analyse des usages de la ville par les migrants dans la Genève des années 1837–1843 [Stay or leave again? An analysis of the use of the city by migrants in Geneva, 1837–1843]. *Annales de Démographie Historique*, 2, 65–87. doi: [10.3917/adh.124.0065](https://doi.org/10.3917/adh.124.0065)
- Remund, A. (2013). Croissance urbaine et durée des épisodes migratoires. L'exemple de Genève au 19ème siècle [Urban growth and the duration of migratory episodes. The example of Geneva in the 19th century]. *Revue Quetelet*, 1(1), 1–17. doi: [10.14428/rqj2013.01.01.02](https://doi.org/10.14428/rqj2013.01.01.02)
- Remund, A. (2014). Des toupies et des enracinés. Mobilité intra-urbaine des immigrés dans la Genève du XIXe siècle [Twirlers and stayers. The intra-urban mobility of immigrants in 19th-century Geneva]. *Annuaire Suisse d'Histoire Économique et Sociale*, 28, 183–203.
- Ritschard, G., & Oris, M. (2005). Life course data in demography and social sciences: Statistical and data-mining approaches. *Advances in Life Course Research*, 10, 283–314. doi: [10.1016/S1040-2608\(05\)10011-2](https://doi.org/10.1016/S1040-2608(05)10011-2)
- Ritschard, G., Studer, M., Müller, N.S., & Gabadinho, A. (2007). Comparing and classifying personal life courses: From time to event methods to sequence analysis. *2nd Symposium of the COST Action 34: Gender and Well-Being*. University of Minho, Guimaraes, Portugal.
- Ritschard, G., Studer, M., & Oris, M. (2009). Analyse statistique implicative des transitions professionnelles dans la Genève du 19e siècle [Implicative statistical analysis of occupational transitions in 19th-century Geneva]. In R. Gras, J.-C. Régnier, C. Marinica & F. Guillet (Eds.), *L'analyse statistique implicative. Méthode exploratoire et confirmatoire à la recherche de causalités* [Implicative statistical analysis. Exploratory and confirmatory method in search of causalities] (pp. 421–435). Toulouse: Cépaduès.
- Ryczkowska, G. (2003). *Accès au mariage et structures de l'alliance à Genève, 1800–1880* [Access to marriage and alliance structures in Geneva, 1800–1880] (Master thesis). University of Geneva.
- Ryczkowska, G. (2013). *Au cœur du social. Le mariage dans le canton de Genève, 1800–1930* [Au coeur du social. Marriage in the canton of Geneva, 1800–1930] (PhD thesis). University of Geneva. doi: [10.13097/archive-ouverte/unige:30736](https://doi.org/10.13097/archive-ouverte/unige:30736)

- Ryczkowska, G., & Perroux, O. (2006). Vieillesse au féminin et au masculin. Individus, familles et collectivité à Genève, 1816–1843 [Ageing for men and women. Individuals, families and community in Geneva, 1816–1843]. *Annales de Démographie Historique*, 2, 189–215. doi: [10.3917/adh.112.0189](https://doi.org/10.3917/adh.112.0189)
- Schumacher, R. (2002). De l'analyse classique à l'analyse différentielle: Nuptialité, fécondité et mortalité à Genève pendant la première moitié du XIXe siècle [From classical to differential analysis: Nuptiality, fertility and mortality in Geneva during the first half of the 19th century] (Master thesis). University of Geneva.
- Schumacher, R. (2010). *Structures et comportements en transition. La reproduction démographique à Genève au 19e siècle* [Structures and behaviour in transition. Demographic reproduction in Geneva in the 19th century] (PhD thesis). Bern: Peter Lang.
- Schumacher, R. (2013). Demographic socialization and reproductive behavior in a transitional context: A macro–micro perspective. *The History of the Family*, 18(2), 154–168. doi: [10.1080/1081602X.2012.712209](https://doi.org/10.1080/1081602X.2012.712209)
- Schumacher, R. (2016). Infant and early childhood mortality in a context of transitional fertility: Geneva 1800–1900. In D. Ramiro Fariñas & M. Oris (Eds.), *New approaches to death in cities during the health transition* (pp. 97–114). Springer International.
- Schumacher, R., Matthijs, K., & Moreels, S. (2013). Migration and reproduction in an urbanizing context. Family life courses in 19th century Antwerp and Geneva. *Revue Quetelet*, 1(1), 19–40. doi: [10.14428/rqj2013.01.01.03](https://doi.org/10.14428/rqj2013.01.01.03)
- Schumacher, R., & Oris, M. (2011). Long-term changes in social mortality differentials. Geneva 1625–2004. *Explorations in Economic History*, 48(3), 357–365. doi: [10.1016/j.eeh.2011.05.011](https://doi.org/10.1016/j.eeh.2011.05.011)
- Schumacher, R., Ryczkowska, G., & Perroux, O. (2007). Unwed mothers in the city. Illegitimate fertility in 19th-century Geneva. *The History of the Family*, 12(3), 189–202. doi: [10.1016/j.hisfam.2007.10.002](https://doi.org/10.1016/j.hisfam.2007.10.002)
- Sewell, W. H. (1985). *Structure and mobility: The men and women of Marseille, 1820–1870*. Cambridge: Cambridge University Press & Paris: Maison des Sciences de l'homme.
- Sommerseth, H. L., & Thorvaldsen, G. (2022). The impact of microdata in Norwegian historiography 1970 to 2020. *Historical Life Course Studies*, 12, 18–41. doi: [10.51964/hlcs11675](https://doi.org/10.51964/hlcs11675)
- Van de Putte, B., & Miles, A. (2005). A social classification scheme for historical occupational data. *Historical Methods*, 38(2), 61–94. doi: [10.3200/HMTS.38.2.61-94](https://doi.org/10.3200/HMTS.38.2.61-94)
- van Leeuwen, M., & Maas, I. (2011). *HISCLASS. A Historical International Social Class Scheme*. Leuven: Leuven University Press.
- van Leeuwen, M., Maas, I., & Miles, A. (2002). *HISCO: Historical International Standard Classification of Occupations*. Leuven: Leuven University Press.
- Vézina, H., & Bournival, J.-S. (2020). An overview of the BALSAC population database. Past developments, current state and future prospects. *Historical Life Course Studies*, 9, 114–129. doi: [10.51964/hlcs9299](https://doi.org/10.51964/hlcs9299)
- Zemon Davis, N. (2015). Geneva, refuge and migrations (16th–17th centuries). Foreword. *Revue de l'histoire des religions*, 1, 5–8. doi: [10.4000/rhr.8340](https://doi.org/10.4000/rhr.8340)
- Zogmal, A. (1998). *Pierre-François Bellot (1776–1836) et le code civil. Conservatisme et innovation dans la législation genevoise de la restauration* [Pierre-François Bellot (1776–1836) and the Civil Code. Conservatism and innovation in Geneva's restoration legislation]. Genève.

HISTORICAL LIFE COURSE STUDIES
VOLUME 12 (2022), published 21-04-2022

Building an Archival Database for Visualizing Historical Networks

A Case for Pre-Modern Korea

Seungmin Paek

Ajou Center for Digital History

Jong Hee Park

Department of Political Science and International Relations, Seoul National University

Sangkuk Lee

Department of History, Ajou University

ABSTRACT

In this paper, we share the experience of collecting and organizing pre-modern Korean historical materials into a searchable digital archive. The Ajou Interdisciplinary Research Group (AIRG) has continuously collected historical data of pre-modern Korea for the past 10 years to assist the study of family history, historical demographics, and social mobility. This paper describes the rich data sources for historical studies of Korea, such as household registers, genealogies, and state examination registers, and we summarize contributions to the study of historical demography and related fields.

Keywords: Korean historical material, Historical demography, Social mobility, Life course studies, Household registers, Genealogies, State examination registers, Longitudinal data, HAVNet DB

DOI article: <https://doi.org/10.51964/hlcs11718>

© 2022, Paek, Park, Lee

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

The Ajou Interdisciplinary Research Group (AIRG) has continuously collected historical data of pre-modern Korea for the past ten years to assist the study of family systems, population, and social mobility. To this end, the AIRG formed a comprehensive research group of experts in history, linguistics, computer science, data science, statistics, and visualization technology. In this paper we introduce the main sources that we describe the AIRG's contributions to historical demography and family history.

Collecting data over a long-time horizon and across a wide range of fields is challenging. More often than not data from different sources lack identifying information, which is necessary to match individuals and a large amount of non-random missingness is very common. Recent advances in data technology in storing, analyzing, and visualizing large-scale data, so-called "big data techniques", provide immense opportunities to easily transform historical data into a digital form. The collaboration of experts in various fields is essential for the digital transformation of historical data, but that alone is not enough. The insights and leadership of historians are essential to lead the entire process of data transformation to an organic workflow. By "organic workflow" we mean that numerous decisions made in the process of data transformation are consistent with each other.

In Korea, government-led research support organizations such as the National Research Foundation and the Korean Academy of Sciences have taken the lead in digitizing historical data since 2000. In this way, the transition of Korean historical data to digital data was carried out on a large scale. Since 2015 the Ajou Interdisciplinary Research Group started seriously using these digitized historical data to establish a database for historical research and social science research. The Historical Archives Visualization Network Database (HAVNet DB) was built to read historical characters and extract information. For a full description of the HAVNet DB see Lee (2016a) and S. Choi, J. Choi, Paek, Yeh, and Lee (2021).

In this paper, we describe the most important sources that form the basis of the HAVNet DB, such as the household registers, the genealogies and the examination lists, and we explain their importance for the study of family history, historical demography and social mobility. Secondly, we give an overview of the results of research in these fields.

2 HISTORICAL MATERIALS OF PRE-MODERN KOREA

Korea is said to be a repository of recorded history. Historical facts have been recorded in various forms for a long span of time by a variety of social actors. The historical materials originate from government as well as from individuals and families. Some of them were listed as UNESCO's Memory of the World Program including *Joseonwangjo sillok* [The Annals of Joseon Dynasty, 朝鮮王朝實錄], *Seungjeongwon Ilgi* [The Diaries of the Royal Secretariat, 承政院日記], *Ilseongnok* [The Records of Daily Reflections, 日省錄], *Uigwe* [The Royal Protocols of the Joseon Dynasty, 儀軌], printing woodblocks of the Tripitaka Koreana and miscellaneous Buddhist scriptures. These various pre-modern historical materials have been digitized in various forms since 2000, making them easily accessible to all interested people. We can classify these sources into unstructured and structured data. Table 1 provides an overview of the websites publishing digitized historical sources, including scans of original sources, tables with structured data, etc. This article focuses on structured data.

2.1 STRUCTURED DATA

2.1.1 HOUSEHOLD REGISTERS

In Korea, *hojeok* [household registers, 戶籍] have been compiled for more than two thousand years. The oldest extant household register is the *Silla jangjeok* [The Silla village document, 新羅帳籍] compiled in the Silla Kingdoms (B.C. 57–935) of which only a small portion remains. The *Silla jangjeok* combined both characteristics of a household register and of a land register. In the Goryeo dynasty (918–1392), the household register and the land register were separated. From the Joseon dynasty (1392–1910) onwards the household registers were compiled every three years for the purpose of taxation and obligatory services, and in principle, all regions and peoples were subject to registration (Son, 2007). The household registers from the Joseon government have survived in some areas, such as Danseong, Daegue, etc. Household registers provide great opportunities to examine demographic behavior over periods of 200 years (about 17th to 20th centuries) by linking individuals and households in consecutive triennial registers into family trees. Moreover, household registers give tremendous opportunities to research and explore social mobility thanks to the recording of people with a variety of social statuses.

Table 1 Websites including Korean historical sources

Data Type	Sources	Websites
Unstructured data	<i>Samguksagi</i> [History of the Three Kingdom], <i>Samgukyusa</i> [Memorabilia of the Three Kingdom], <i>Goryeosa</i> [History of Goryeo Dynasty], <i>Goryeosa jeoryo</i> [Condensed History of Goryeo Dynasty], <i>Joseonwangjo sillok</i> [The Annals of Joseon Dynasty], <i>Seungjeongwon Ilgi</i> [The Diaries of the Royal Secretariat], <i>Ilseongnok</i> [Records of Daily Reflections], etc.	Korean History Database by National History Compilation Committee
	<i>Uigwe</i> [The Royal Protocols of the Joseon Dynasty], Historical documents, Stone and bronze inscriptions, etc.	Kyujanggak Institute for Korean Studies, Digital Archives for Korean Studies Digital Jangseogak
Structured data	Household Registers of Danseong and Daegu-bu in Joseon Dynasty	Human Resources Information System in Korean History Daedong Institute for Korean Studies
	Jokbo [genealogy]	Korean Jokbo Data System
	The List of State Examination Rosters, etc.	Korea Historical Information Integration System

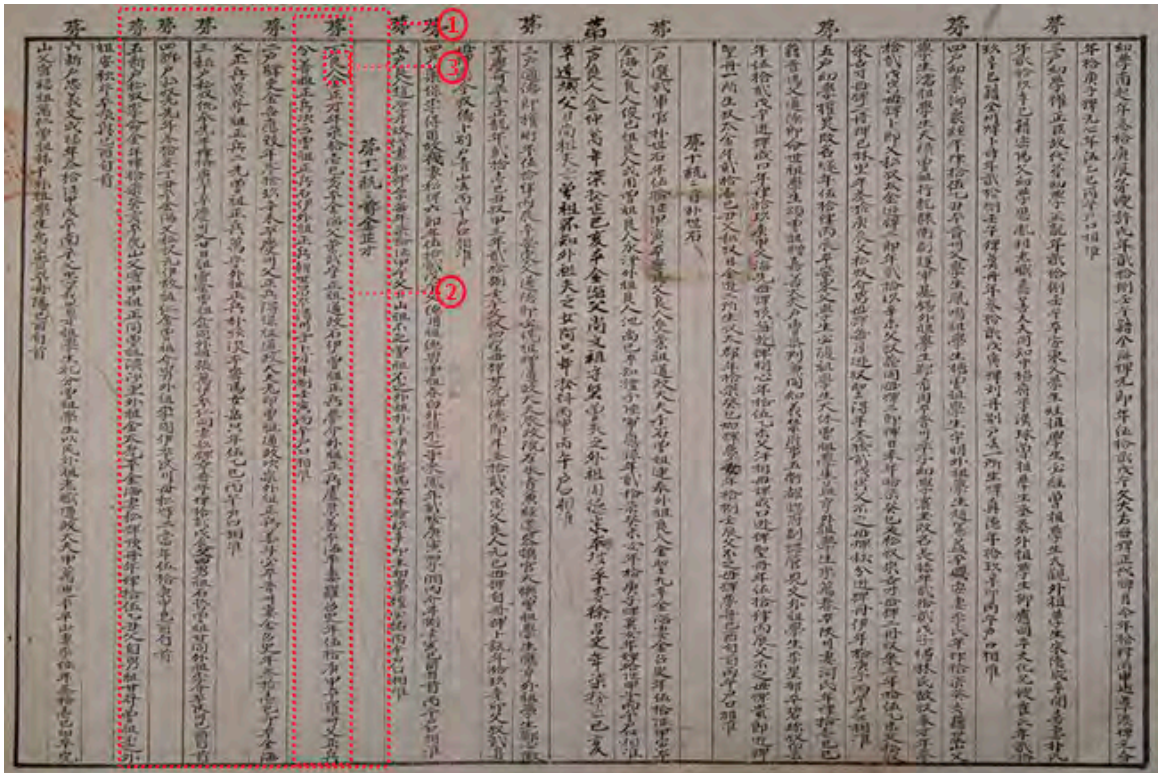
From all Joseon dynasty household registers, those from Danseong and Daegu (see map 1) are the most well-known in the academic world. Danseong household registers recorded people who lived in the Danseong area (present Sancheong-gun county, Gyeongsangnam-do province) through 1606–1888. The Danseong area was a typical rural area. On the other hand, Daegu (present Daegu Metropolitan City) was a mixed area with urban and rural characteristics. Daegu was one of the major administrative cities of the Joseon dynasty, where the headquarters of the provincial governor [*gamyeong*, 監營] of the Gyeongsang-do province was established. After 1601, Daegu-bu expanded its size as it developed into an administrative hub of the Gyeongsang-do province. In addition, Daegu was also one of the top three commercial cities of the late Joseon dynasty. In other words, Daegu-bu was not only a representative administrative city but also served as a center of commerce. Accordingly, *seosang-myeon* and *dongsang-myeon*, where *gamyeong* was located, were areas where urban characteristics were strongly reflected, while the rest of the areas were typical rural areas. Besides those of Danseong and Daegu, also the household registers in Ulsan, Eonyang, Jeju, and Sangju have survived.

Map 1 South Korea, with the areas where Joseon's household registers [*hojeok*] have survived



In the following we will explain what elements were recorded in the household registers. Figure 1 shows a copy of an original page of the Danseong household registers. These registers are organized in *tongs* each composed of five households, one *li* [village, 里] consists of dozens of *tongs*. Several *lis* make up one *myeon* [township, 面]. The Danseong household registers include eight *myeons* of the county of Danseong-hyeon. In the middle of the example in Figure 1, we can identify the 11th *tong* of hyunnae-myeon township, Danseong-hyeon county in 1717. The 11th *tong* is composed of five *hos* [household, 戶] (①). Each household is recorded with a head, his wife, their children, their siblings, parents, grandfathers, great-grandfathers, and fathers-in-law of a head and his wife, and *nobi* [unfree people, 奴婢] in terms of guidance of the National Code of Joseon Dynasty [gyeongguk-daejeon, 經國大典] (promulgated in 1485) (②). Each person in a household has their *jikyeok* [occupational title, 職役] such as *seonmugungwan* [elected military officials, 選武軍官], *yangin* [commoner, 良人], and *nobi*, etc. (③). Unlike the Chinese household registers of the northeast Chinese province of Liaoning (Lee & Campbell, 2016), numbering of the *tong* and the *ho* was variable and could be changed in the next consecutive compilation year. Even members and their *jikyeoks* in a *ho* might be changed in the next consecutive year. Therefore, we can examine changes in their social status from a longitudinal perspective if we can link individuals and household data into life courses. The Daedong Institute of Sungkyunkwan University transcribed and converted the texts of Danseong and Daegu household registers into a data structure for this kind of research.

Figure 1 Example of the Danseong household registers



Note: For the explanation of the reference numbers, see the text.

Figure 2 A capture of the digitized Danseong household registers

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
year	myeon	no	li_no	li	tong	t_head	ho_no	hh_head	rs_to_hhh	occ_title	s_name	g_name	age	sexa_cyc	exit	place	new
1825	sindeung	1	1	dangyesang	1	統管	1	jeongdoli	head	hanyang	jeong	doli	78				
1825	sindeung	2	1	dangyesang	1		1	jeongdoli	wife		son	sosa	68				
1825	sindeung	3	1	dangyesang	1		2	simseokyeong	head		sim	seokyeong	31				
1825	sindeung	4	1	dangyesang	1		2	simseokyeong	mother-in-law		kwon	ssi	52				
1825	sindeung	5	1	dangyesang	1		2	simseokyeong	wife		park	ssi	35				
1825	sindeung	6	1	dangyesang	1		2	simseokyeong	brother			gyouyeong	14				
1825	sindeung	7	1	dangyesang	1		2	simseokyeong	nobi			jaeran	52				branch family
1825	sindeung	8	1	dangyesang	1		2	simseokyeong	nobi			woonjeol	31				
1825	sindeung	9	1	dangyesang	1		2	simseokyeong	nobi			gabsam	32				
1825	sindeung	10	1	dangyesang	1		2	simseokyeong	nobi			boksam	52				
bon	anc_seat	ow_place	ow_occ	ow_name	f_occ	f_name	m_occ	m_name	bir_order	gt_occ	gt_name	ggf_occ	ggf_name	mg_occ	mg_name	mg_anc_seat	memo
bon	gyeongju					seonbong					buji						
											buji						
bon	cheongsong					yuhak					haksaeang	saryang	tongdeokrang	Jeongsin	haksaeang	kwonjoonhae	andong
jeok	andong																
jeok	goryeong										haksaeang	sanggon	haksaeang	don	haksaeang	hwangjaehae	jangsoo

Note: For the explanation of the reference numbers, see the text.

Table 2 Basic statistics of the households register datasets from the Danseong and Daegu areas

	Time span	Myeon [面, township]	Household records	Person records
Danseong	1606–1888	8	42,795	240,692
Daegu	1681–1876	30	275,073	1,516,273
Total		38	317,868	1,756,965

Source: HAVnet DB, see Choi et al. (2021).

Figure 2 shows a screenshot of the digitized household registers in Excel format. For each individual, the following information is available: the name of household head (①), relationship to household head (②), occupational title (③), surname and given name (④), age calculated by subtraction from compilation year minus year of birth (⑤), year of birth (in animal sign with the period name from the sexagenarian cycle) (⑥), exit from the register (emigration, out-marriage, death, escape) (⑦). Other fields (not shown in figure 2) describe the entry into the register, ancestral seats, the occupational titles and names of parents, grandfathers, great-grandfather, and fathers-in-law of a head and his wife. At the end of each *myeon* [township], informative statistics of each *myeon* are recorded, and at the end of each year of the register, various official statistics constructed by the local government such as the number of households, so called *doisang* [都已上], are recorded. This official statistic from *doisang* does not match the actual statistic calculated from the text of the household registers because the goal was not to show the actual number of individuals and households of the area, but to secure the necessary tax sources for the government. The range of demographic information contained in Danseong household registers is very rich and can be used for the study of families over a long time (Park & Lee, 2008). Table 2 shows the total number of households and persons included in the dataset of the household registers, adding up to about 1.8 million person records from 38 townships of Danseong and Daegu household registers covering the period 1606–1888.

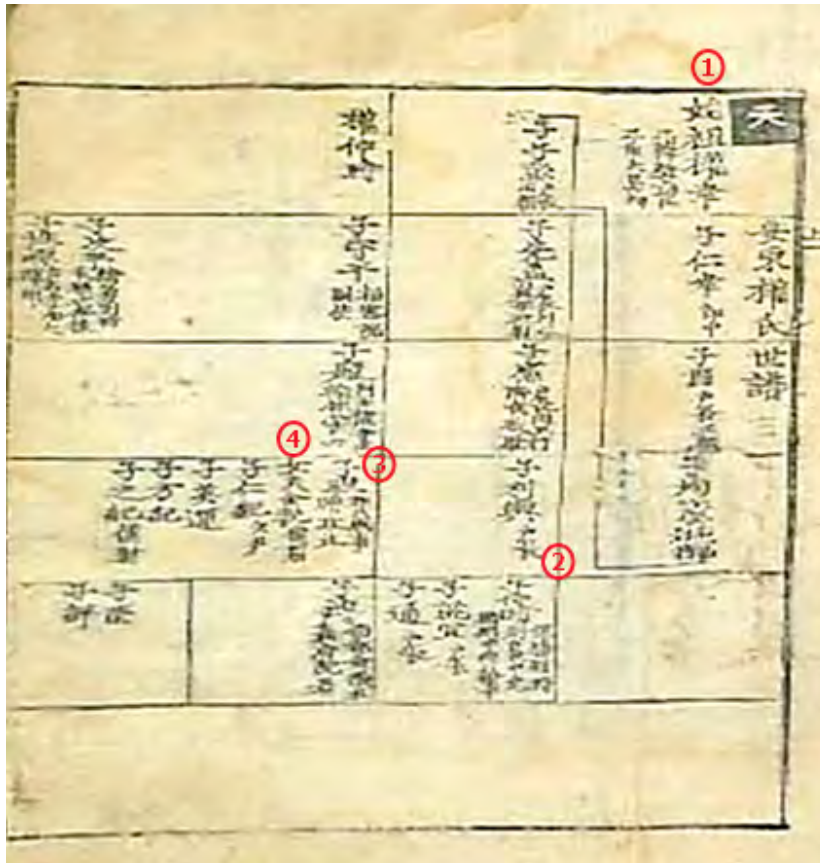
2.1.2 FAMILY GENEALOGY

The *jokbo* [genealogy, 族譜] consists of multi-generational records of families from the progenitor of the clan to his descendants at the time of publication. It has been published continually for hundreds of years from the early Joseon dynasty and it is still being produced to this day. A clan in *jokbo* refers to the paternal blood relationship, which honors the same ancestor by carrying the same surname [姓] and ancestral seat [本貫]. Although there are differences depending on the period and type of publication, the Korean genealogy generally records individual demographic records with date and place of events as birth, death and official ranks.

Andong Gwon-ssi clan's *jokbo*, called *Seonghwa-bo* [成化譜], is the oldest extant genealogy in Korea, which was first published in 1476. *Seonghwa-bo* contains detailed demographic information of more than 10,000 Andong Gwon-ssi clan members. According to *Seonghwa-bo*, the progenitor of the Andong Gwon-ssi clan is Gwon, Haeng, who lived in the early Goryeo dynasty (918–1392). The last entry of *Seonghwa-bo* is the 21st generation of the progenitor. Another important *jokbo* in early stage is *Gajeong-bo* [嘉靖譜] of the Munhwa Ryu-ssi clan. *Gajeong-bo* was first published in 1565 and contains detailed demographic information of more than 49,000 Munhwa Ryu-ssi clan members. The progenitor of Munhwa Ryu-ssi clan is Ryu, Chadal and the last entry of *Gajeong-bo* is the 24th generation. The publication of these two *jokbos* by powerful elite families of the Andong Kwon-ssi clan and the Munhwa Ryu-ssi clan prompted the publication of *jokbo* by other elite families in the 16th century. The number of genealogical records published before 1945 is known to be 3,389 *jokbos* for 143 different surnames (Lee and Park, 2008).

Figure 3 is a picture of the first page of *Seonghwa-bo*, showing the beginning of the Andong Gwon-ssi clan from the progenitor to the 13th generation. The first person listed is the progenitor, Gwon, Haeng [權幸] (①) who received an official position as a vassal of merit from Wang, Geon, the founder of the Goryeo dynasty. Before the 9th generation, only one member was recorded per generation. From the 9th generation onwards, also a second sibling began to be recorded (②). At the 12th generation (③), six siblings (five sons and one daughter) were recorded by birth order. Note that the recording by birth order, instead of gender, is unique to these early family genealogies. After the 17th century, the recording order changed into putting the sons first and the daughter at the end of the row, reflecting social changes.

Figure 3 First page of the Andong Gwon-ssi Clan's Jokbo (Lee & Lee, 2017)

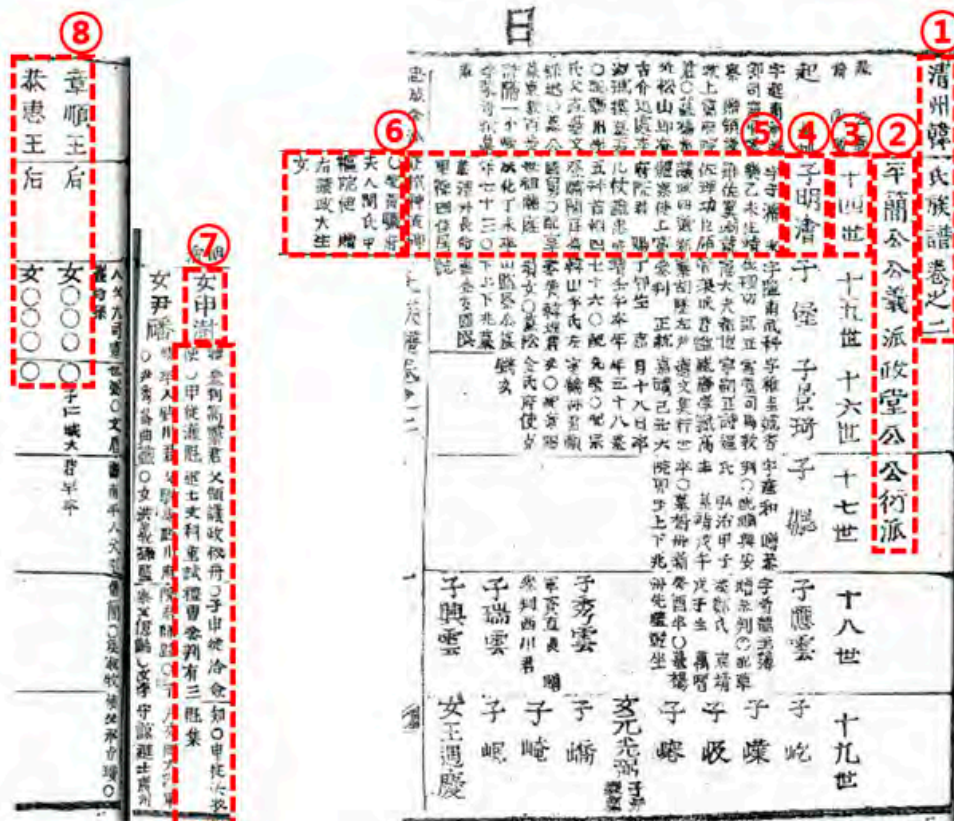


Note: For the explanation of the reference numbers, see the text. The source has to be read from right to left and from top to bottom, each rectangle presents one generational level.

Another distinct aspect of Korea's family genealogy is that *jokbo* records not only the complete list of sons' family lines but also that of the daughter's family lines. When a daughter marries, the record of her husband is listed in her family's *jokbo* as a son-in-law [*yeobu*, 女夫] (④) of her family's head. The Korean concept of family embraces in-laws as family members. For this reason, the Korean genealogy contains a wide range of information not only about members of the family, but also about members of other prestigious families outside the family. This makes Korea's family genealogies a valuable resource for the study of the elite society in pre-modern Korea (Lee & Lee, 2017).

The *jokbo* is abundant in the demographic, historical, and sociological information on family members. Figure 4 shows another example, which is the second volume of the *jokbo* of the Cheongju Han-ssi clan in late Joseon dynasty (①), presenting the 14th and further generations. The Cheongju Han-ssi clan branched out into multiple lines, called "pa" [派], two of which are represented in this *jokbo*: *Pyeonggangong Gongui-pa* [平簡公 公儀派] and *Jeongdangong Gongyeon-pa* [政堂公 公衍派] (②) which indicate the first name of the person and his government office title, respectively. If we move to the 14th generation [十四世] line (③), we can find the name of *Myeonghoi* [明澮] (④), who was the first-born son (子) of the previous generation. More information can be found from ⑤ such as the courtesy name [字], given to a man when he becomes an adult expressing specific preferences or virtues, the year of birth, the career as government official, the spouse and information about the affinal family, the year of death, and the location of his grave. Reading this line of information, we can also find information about the husbands of the daughters of *Myeonghoi* [明澮] (⑦). For example, *Myeonghoi's* first son-in-law is *Sin, Ju* [申澗], his government office was *Jeong Champan* [贈 參判], and his government office title was *Goryeong-gun* [高靈君]. Moreover, we can also find that his father's name was *Sin, Sukju* [申叔舟] reaching *Yeonguijeong* [the chief state counselor, 領議政], which is one of the most prestigious government positions. *Myeonghoi's* two daughters were queen consorts. Instead of a queen's name, the two daughters' posthumous titles are reported as *Jangsun-wanghu* [Queen Jangsun, 章順王后] and *Gonghye-wanghu* [Queen Gonghye, 恭惠王后], which can be interpreted as 'the gentle queen' and 'the benevolent queen' respectively (⑧). The two daughters' titles are written in a new line which means that they were honored with a lot of respect as being members of the royal family members (Hong, Lee & Yoo, 2021).

Figure 4 A part of Cheongju Han-ssi Sebo in late Joseon dynasty (Hong, Lee, & Yoo, 2021)



Note: For the explanation of the reference numbers, see the text. The source has to be read from right to left.

Figure 5 Example of the digitization of the Andong Gwon-ssi Clan's Jokbo

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
code_UWJ	last_name	last_name	first_name	string	line	sex	son_in_law	Gov_Exam	op_rank	op	ancestral_seat	period	birth_year	death_year	op_year	fa_name	grandfa_n	greatgrani	greatgreatgrandfa_n	
1	kwon	1	haeng	a 1	0	0	0		1	삼한백성공신	andong									
2	kwon	1	inhaeng	a 1 1	0	0	0		5	낭중	andong									
3	kwon	1	chaek	a 1 1 1	0	0	0		12	홍장정조	andong									
4	kwon	1	gyunhan	a 1 1 1 1	0	0	0		12	홍장정조	andong									
5	kwon	1	japaeng	a 1 1 1 1 1	0	0	0		12	홍장정조	andong									
6	kwon	1	seongae	a 1 1 1 1 1 1	0	0	0		12	홍장정조 익아교위	andong									
7	kwon	1	ryeom	a 1 1 1 1 1 1 1	0	0	0		12	홍장정조 행배음교위	andong									
8	kwon	1	lyeo	a 1 1 1 1 1 1 1 1	0	0	0		12	홍장	andong									
9	kwon	1	joangsi	a 1 1 1 1 1 1 1 1 1	0	0	0		12	보승별장 부포장중윤	andong									
10199	kwon	1	chilyi	a 1 1 1 1 1 1 1 1 1 2	0	0	0		12	포장	andong									
10203	kwon	1	tong	a 1 1 1 1 1 1 1 1 3	0	0	0		12	포장	andong									
10	kwon	1	supyeong	a 1 1 1 1 1 1 1 1 1 1	0	0	0		3	추밀원부사	andong				1250					
6061	kwon	1	chapyeong	a 1 1 1 1 1 1 1 1 1 2	0	0	0		13	정충보장 식육정	andong									
6062	kwon	1	seongwon	a 1 1 1 1 1 1 1 1 1 3	0	0	0		14	대연사	andong									

Note: For the explanation of the reference numbers, see the text.

We are in the process of digitizing family genealogy information in the format shown in Figure 5. The principle of digitization is to ensure that all the relationships between the individuals can be easily identified. To this end, we have developed a simple, precise, and easy-to-understand coding scheme for family genealogy. This coding scheme (①) can be explained in the following algorithm.

- The progenitor of a family genealogy is recorded as 1. In Figure 5, Gwon, Haeng is the progenitor and he is recorded as 1.
- The offspring is recorded by their birth order. In Figure 5, Gwon, In-haeng is the first-born of Gwon, Haeng and hence recorded as 11.

Then, the length of the digit indicates the generation and the last digit indicates the ego's birth order (starting with the 9th generation in Figure 5). Also, we can easily track the ancestry history by comparing the sequence of the digit. For example, 1114 indicates that the ego is the fourth child in the 4th generation from the progenitor and the fore-going generations are all first-borns (great-grandfather, grandfather, and father). The ego's siblings can be identified by the same ancestor history (111). In this way, we can easily identify relatives of the ego. For example, nieces and nephews in the 4th generation are indicated by having the same first two digits: 1121, 1134, etc.).

For the title of the official rank in the governmental bureaucracy recorded in the genealogy, we use various historical materials to find the most accurate information of individuals' titles of the official ranks. Fortunately, the government offices are clearly divided into nine levels [poom]. Thus, the official ranks are coded from 1 (highest) to 9 (lowest) (②). We also coded for ranks outside the government offices: 12 and 13 for a provincial official, 14 for a Buddhist monk, 15 for royal family members, and so on. A more daunting task in collecting office information is how to match records from different sources. We first use the UCI which stands for the Universal Content Identifier developed by the Academy of Korean Studies (<http://people.aks.ac.kr/front/uci/uciInfo.aks?isEQ=false&kristalSearchArea=P>). The basic structure of UCI consists of three parts:

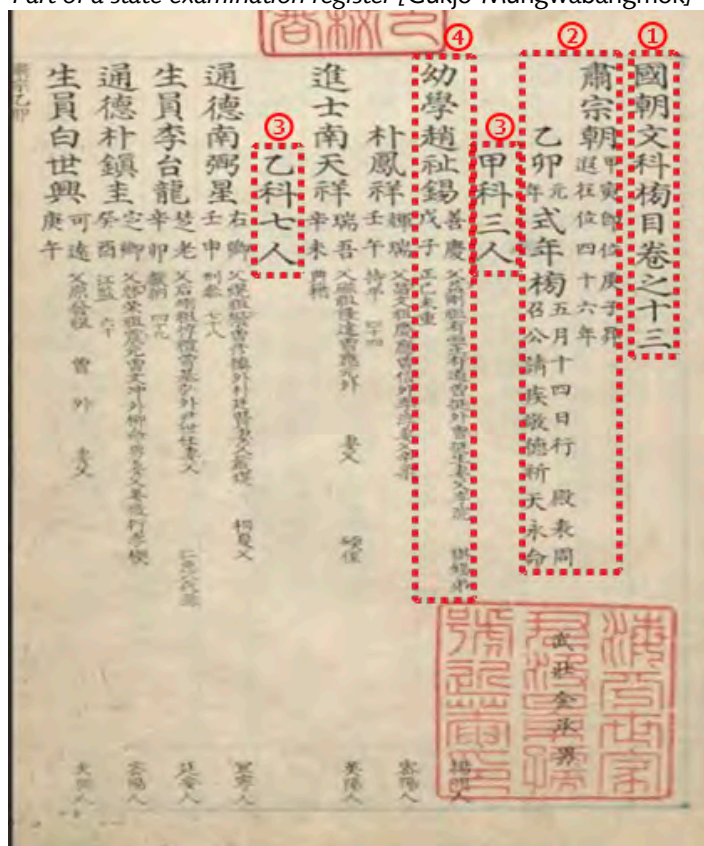
- institution code;
- code to classify the source of the historical data;
- a unique individual identifier: gender, name by Unicode encoding, birth year, death year.

For documents without a UCI, we use comparable identifying information to match individuals across documents, namely the first name, family name (both in Chinese characters), birth year, death year, and father's name.

2.1.3 STATE EXAMINATION REGISTERS

Persons who successfully passed the civil service examinations during the Joseon dynasty were compiled in a list, called *Gukjo-Mungwabangmok* [國朝文科榜目]. The list contains the entire record of 15,151 successful candidates of the civil service examination for the 804 examinations from 1393 to 1894. Among all the state examinations, the civil service examination [*mungwa*, 文科] was considered to be the most important for an individual to acquire a public office. The Academy of Korean Studies collects various individual-level information on 15,151 successful candidates and makes them publicly available (<http://people.aks.ac.kr/index.aks>).¹

Figure 6 Part of a state examination register [*Gukjo-Mungwabangmok*]



Note: For the explanation of the reference numbers, see the text.

1 Because only successful applicants are recorded, information on the failed applicants is not known. But it is possible to estimate the number of people that applied looking at historical materials such as The Annals of Joseon Dynasty.

Figure 6 shows the basic structure of *Gukjo-Mungwabangmok*. The beginning of the page shows the title and volume number of the book (①). In the column referenced by ②, we recognize when the examination was carried out — in this case 1675, the first year of King Sukjong's reign — and what type of the examination it is [*singnyeonsi*, 式年試, regular examination]. We know the examination division and ranking of the individual out of total number of successful candidates: *Gabgwa* [甲科, first-grade group] is successful candidates ranked in 1st, 2nd, 3rd, followed by *eulgwa* [乙科, second-grade group] for 4th through 7th ranks, and the last one is *byeonggwa* [丙科, third-grade group] in ③. The next column ④ presents a variety of individual information on the successful graduate including their names and offices, previous office positions, and the name of their fathers, grandfathers, great-grandfathers, fathers-in-law, and grandfathers-in-law. What is interesting in *Gukjo-Mungwabangmok* is that the document records the names of all the successful candidates' family members who also passed the exam, regardless of their blood ties. That is, the *Gukjo-Mungwabangmok* registers also the names and office positions of the successful candidate's maternal as well as paternal ancestors who were also successfully examined.

Figure 7 shows a part of a table with data compiled from the examination lists. The dataset can be divided into three types of data: individual demographics, family-related information, and career-related information. The demographic data includes the surname (①), first name (②), birth and death year (③) and region of living (④). The family-related information includes the ancestral seat [*bongwan*, 本貫] and the names and office records of the father. Comparable information from the grandfather, great-grandfather, father(s)-in-law and grandfather(s)-in-law is also included but not shown in figure 7. The career-related information includes the entire history of the office records that were retrieved from The Annals of Joseon Dynasty [*Joseonwangjo sillok*, 朝鮮王朝實錄]. Because of the wide range of information at the individual and family level, the data provides tremendous potential for various purposes of research. For example, the data can be used to study how family background of an individual affected his social mobility, how an individual's exam performance shaped the history of his career over his lifetime, or the process a family lineage used to successfully produce more of their members in high offices and become more powerful.

Table 3 summarizes the number of examinations and successful candidates for each king's reign. Note that there are two types of civil service examination: regular and irregular examinations. Regular examinations were held every three years and account for 39.8% of all cases and irregular examinations account for the rest. The number of examinations and the number of successful candidates increase over time because of the more frequent implementation of irregular examinations. Irregular examinations were conducted without notice when the state had something to celebrate such as the inauguration of a new king, the birth of a prince, the kings journey to the provinces, etc.

Figure 7 Screenshot of a table with state examination graduates

	A	①	②	D	E	F	G	H	③	④	L	M	N	O
1	순번	성	명	성명	자	호1	본관	생년	몰년	거주지	급제시험종류	급제식년	급제년도	합격등급
2	1	송	개신	宋介臣	미상	휴재(休齋)	홍주(洪州)	1373	미상	미상(未詳)	식년시(式年試)	태조2	1393	을과(乙科) 1[壯元]위
3	2	김	호원	金孝源	미상	미상	김해(金海)	1370	미상	미상(未詳)	식년시(式年試)	태조2	1393	을과(乙科) 2[亞元]위
4	3	이	담	李擔	미상	미상	경주(慶州)	1370	1405	미상(未詳)	식년시(式年試)	태조2	1393	을과(乙科) 3[探花郎]위
5	4	탁	함	卓咸	미상	미상	광산(光山)	1345	미상	미상(未詳)	식년시(式年試)	태조2	1393	병과(丙科) 1위
6	5	윤	정	尹定	미상	미상	함안(咸安)	1376	미상	미상(未詳)	식년시(式年試)	태조2	1393	병과(丙科) 2위
7	6	변	계순	卞季孫	미상	미상	조계(草溪)	1368	미상	미상(未詳)	식년시(式年試)	태조2	1393	병과(丙科) 3위
8	7	변	처후	邊處厚	미상	미상	장연(長淵)	1373	1437	미상(未詳)	식년시(式年試)	태조2	1393	병과(丙科) 4위
9	8	신	개	申龜	자격(子格)	인재(仁齋)	평산(平山)	1374	1446	미상(未詳)	식년시(式年試)	태조2	1393	병과(丙科) 5위
10	9	홍	중강	洪仲剛	미상	미상	남양(南陽[唐])	1373	미상	미상(未詳)	식년시(式年試)	태조2	1393	병과(丙科) 6위
11	10	이	숙번	李叔蕃	미상	미상	안성(安城)	1373	1440	미상(未詳)	식년시(式年試)	태조2	1393	병과(丙科) 7위
12	11	소	호인	蘇好仁	미상	미상	진주(晉州)	1356	미상	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 1위
13	12	유	흘	柳洽	미상	미상	백천(白川)	1356	미상	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 2위
14	13	김	식	金湜	미상	미상	미상(未詳)	1356	미상	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 3위
15	14	서	순	徐選	미상	미상	이천(利川)	1367	1433	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 4위
16	15	김	호손	金孝孫	미상	미상	의성(義城)	1373	1429	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 5위
17	16	이	조	李椒	미상	미상	광산(光山)	1372	미상	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 6위
18	17	한	겸	韓兼	미상	미상	평산(平山)	1365	미상	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 7위
19	18	송	호	宋瑚	미상	미상	태인(泰仁)	1373	미상	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 8위
20	19	유	의	柳儀	미상	미상	고흥(高興)	1373	미상	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 9위
21	20	나	득경	羅得卿	미상	미상	나주(羅州)	1369	미상	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 10위
22	21	최	예	崔潏	미상	미상	경주(慶州)	1371	미상	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 11위
23	22	민	안	閔顔	미상	미상	여흥(驪興)	1350	미상	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 12위
24	23	박	조	朴漵	미상	미상	미상(未詳)	미상	미상	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 13위
25	24	한	련	韓璉	미상	미상	미상(未詳)	1367	미상	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 14위
26	25	황	현	黃鉉	미상	미상	평해(平海)	1372	미상	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 15위
27	26	이	증명	李仲明	미상	미상	미상(未詳)	1370	미상	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 16위
28	27	김	자린	自麟	미상	미상	광산(光山)	1372	미상	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 17위
29	28	이	관	李灌	미상	미상	인천(仁川)	1372	1418	미상(未詳)	식년시(式年試)	태조2	1393	동진사(同進士) 18위

Note: For the explanation of the reference numbers, see the text.

Table 3 Number of exams and successful examinees of the Civil Service Examination, 1393–1894

King (Year)	Exams	Successful Graduates
Taejo (1393–1396)	2	66
Jeongjong (1399–1399)	1	33
Taejong (1401–1417)	11	281
Sejong (1419–1447)	21	510
Munjong (1450–1451)	2	73
Danjong (1453–1454)	3	106
Sejo (1456–1468)	23	402
Yejong (1469–1469)	1	33
Seongjong (1470–1494)	29	473
Yeonsan (1495–1506)	13	261
Jungjong (1506–1544)	57	933
Myeongjong (1546–1566)	26	497
Seonjo (1567–1606)	61	1,129
Gwanghae (1608–1621)	28	510
Injo (1623–1649)	52	784
Hyojong (1650–1657)	15	253
Hyeonjong (1660–1673)	24	405
Sukjong (1675–1719)	78	1,465
Gyeongjong (1721–1723)	9	183
Yeongjong (1725–1776)	126	2,179
Jeongjo (1776–1800)	41	798
Sunjo (1801–1834)	51	1,058
Heonjong (1835–1849)	23	463
Cheoljong (1850–1863)	26	476
Gojong (1864–1894)	81	1,780
Total	804	15,151

Source: <http://people.aks.ac.kr/index.aks>

2.2 UNSTRUCTURED DATA

There are many sources with unstructured data written by the government as well as individuals and families in Korean history. As explained above, some of them were listed as UNESCO's Memory of the World Program. Out of them, we introduce two representative historical materials by the Joseon government. First, the *Goryeosa* [高麗史, History of Goryeo Dynasty] is a historical book compiled by Jeong, In-ji and others who were the Confucian scholars under order of King Taejo (reign year 1392–1398, founder of Joseon dynasty). The *Goryeosa* is considered as one of the most authoritative sources for studying the Goryeo dynasty (918–1392). The *Goryeosa* was known to be compiled in 1449 since King Taejo ordered, completed in two and a half years, and finally published in 1454. The composition of the book follows the annalistic form [紀傳體], consisting of 139 volumes. Among them, 46 volumes are annals of kings [sega, 世家], 39 volumes are brief descriptions of important facts, systems, geography, etc. [ji, 志], 2 volumes are time tables [pyo, 表], and 50 volumes are arrayed biographies of historical figures [yeoljeon, 列傳].

A second example of unstructured data is The Annals of Joseon Dynasty [*Joseonwangjo sillok*, 朝鮮王朝實錄]. These annals consists of official records of Joseon kings over 25 generations and 472 years from 1392

onwards.² They are considered as the most important resource to study the history and culture of the Joseon dynasty. The Annals of the Joseon Dynasty are recorded in chronological form [編年體] and totaling 1,893 volumes. They were usually published during the period of a king's succession. The Annals of the Joseon Dynasty are considered as valuable historical records that contain detailed history of various aspects of Joseon including politics, diplomacy, military, system, law, economy, industry, transportation, communications, society, customs, astronomy, geography, science, medicine, literature, music, art, crafts, ideology, ethics, religion, etc. Figure 8 shows a screenshot of the webpage that provides the access to the book. The website currently provides digitized contents in Chinese classical characters and in Korean language. The row indicated with ① gives the date that the specific historical event happened and ② supplies the abstract of the event in Korean. A full description of the event appears both in Korean and Chinese classical characters (③). And the row indicated with ④ presents the original references from The Annals of Joseon Dynasty.

Figure 8 Capture of The Annals of the Joseon Dynasty from Korean History Database by National History Compilation Committee



Note: For the explanation of the reference numbers, see the text.

3 RESEARCH ON DEMOGRAPHIC BEHAVIOR, FAMILY HISTORY, AND SOCIAL MOBILITY IN KOREA

In this section, we highlight research results for demographic behavior, family history, and social mobility by the Ajou Interdisciplinary Research Group (AIRG) and other researchers. These contributions to social science history reflect the dedicated work of many Korean scholars. As we review these studies, we look back on previous work and establish milestones for future studies.

2 The Annals of the last two kings including King Gojong and King Sunjong of the Joseon dynasty are not recognized as The Annals of the Joseon Dynasty because they were published during the Japanese colonial period. Refer to <http://esillok.history.go.kr/about/veritableRecordsInfo.do?sessionId=9B8A53FBBBD024C8A14FE3E4CD55395E>.

3.1 DEMOGRAPHIC STUDIES

Korea's rich household registers and family genealogies have long been considered valuable resources for historical demographic research (Wagner, 1974a). Social demographers used these data to understand the population dynamics of the Joseon dynasty (Cha, 2009; Kwon & Shin, 1977). A major breakthrough came with the digitization of the Danseong and Daegu household registers since around 2000, which allowed full-fledged historical demographic research using computerized data. However, after carefully reviewing the nature of the collected data, these scholars concluded that both the household registers and the genealogical data were recorded by the state or family in a selective way and hence these data could not be considered as a representative sample of the population (Household Registers Research Team, 2003; Lee, 2010a; Park & Lee, 2008).

These findings have had a significant impact on future research of historical demography in Korea. First, scholars realized that the history of the entire population of the Joseon dynasty could not be understood using the household registers and genealogical data. They record only a part of the historical reality from which we do not know how to construct missing demographic data (Jung, 2007; Kim, Park, & Jo, 2013). Scholars warn that strict verification is necessary to use the genealogies and the household registers as data for historical demographic studies (Miyajima, 2004; Son, 2016). Second, as a result, scholars turn their attentions to comparative historical studies in which characteristics of Korean data are compared with those of the West and China (Han, 2020; Park & Lee, 2008; Rhee, 2004).

3.1.1 FERTILITY AND MORTALITY

Thanks to studies that are a milestone in data utilization, Korea has actively contributed to historical demographic research. Above all, using genealogical and household register information on birth and death, researchers have estimated life expectancies and fertility rates. Male life expectancy at birth was estimated to be 23 years during the 18th and 19th centuries, based on information about mortality in the early 20th century and model life tables from genealogies. Age-specific marital fertility rates for upper class females were calculated from genealogies and were combined with estimates of age at first marriage and information on colonial fertility to derive a total fertility rate of 6.81 (Cha, 2009). Given that life expectancy of the nobility during the Goryeo dynasty (918–1392) was 34.8 years (Lee, 2010a), the gap between the figures from this period and the 18th and 19th centuries was very large. More case studies are needed to arrive at more definitive conclusions, because the results vary depending on the circumstances of the data.

It has been confirmed that fertility is affected by social status and economic power. A study on the household registers of Jeju Island of Korea from 1914 to 1925 found a positive relationship between the size of the land holdings and the childbearing. Since this relationship was not linear, an analysis of nonlinear relationships was attempted using qualitative examples (Kim & Park, 2009). Marital fertility was also investigated through data from the family registers of the Japanese colonial period. This study suggested that improvement in child survival is a prerequisite for lower birth rates (Kye & Park, 2016). Women's age at first childbirth was estimated using household registers of late 17th to early 18th century Joseon dynasty. Family histories reconstructed by the connection of consecutive household registers to compensate for the defects of these registers confirmed that cultural factors such as social status and marriage customs are closely related to fertility (Son & Lee, 2010). Combining household registers and genealogical data from the 19th to the mid-20th century made it possible to determine the relationship between the social status of parents and the number of children. High socioeconomic status proved to be a factor in having more children (Lee & Yoo, 2018).

Research has also been focused on mortality. The nobility belonging to the Goryeo dynasty (918–1392) shows a lower level of mortality than the nobility in China and England during the same period (Lee, 2010a). This result is limited to an elite population, but it is important to learn that mortality was lower in Korea than in China and England at the time. Estimates of child mortality in mid-20th-century population registers have examined the effects of gender, birth order and sibling composition. The relative strength of social and biological factors on mortality was by historical context, and Korean families actively respond to these constraints (Park, Han, & Kye, 2018).

3.1.2 MARRIAGE, ADOPTION, AND MIGRATION

Male and female ages at marriage have been calculated over the period 1678–1789 using Danseong household registers. Most first marriages took place between the ages of 15 and 20. During this period, marriage in Korea was earlier than in Europe, later than in Japan, and similar to China (Kim, 2005). Marriages tended to be homogeneous in terms of social background. Since the 17th century, when tribal villages

organized on the basis of patriarchal blood relations [*dongseong chonrak*, 同姓村落] gradually became more concrete, marriage to spouses of other influential family members living in nearby areas has become common (Kwon, 2006). Comparisons of urban and rural marriage have been conducted using the household registers of Daegu in the 18th century. Differences in marriage ages and remarriage rates are attributed to differences between agricultural and urban lifestyles, value systems, and occupations (Kim, 2009).

Studies of family succession have considered duration of residence, adoption, and re-marriage. The specific characteristics of Korean adoption differ from practices in China and Japan (Kim & Park, 2010). Upper-status genealogies [*Bulcheonwye jokbo*, 不遷位族譜] suggest that the close relationship between adoption and birth rates is only partly explained by cultural factors such as status maintenance. The correlation between male remarriage and adoption has also been examined. The number of remarriages among men peaked at the end of the 17th century and decreased rapidly until the 19th century. Meanwhile, adoption was more frequently used to maintain the social and economic status of the family (Son, 2010).

A recent genealogical analysis of Korea's 13th- to 15th-century marriage networks examines the role of marriage in family succession and maintenance of social status (Lee & Lee, 2017). Marriage was an important strategy of maintaining social status. Even if the existing power structure changed because a dynasty was replaced, political elites wanted to maintain the existing marriage network. However, appointments of new officials outside the existing family network showed that the system was not as closed as one might think. Along the same line, the impact of marrying into the royal family was also studied (Hong, Lee, & Yoo, 2021). Genealogies of 15 elite families show that they regarded marriage as a means of managing the socio-political inner circle of elite families in early Joseon Korea (1392–1506). Marriage patterns indicate that the socio-political power of affinal kin has a greater effect on promotions than descent or meritocratic considerations. In particular, marrying into a queen consort's family increased the likelihood an individual would end up in a high position, which was beneficial for retaining the political power of the family.

Although migration is a significant life course event, it is not studied as often as it should be. Nevertheless, several studies of migration have used household registers from the 19th century to the Japanese colonial period in Korea (1910–1945). One study examined patterns of geographic mobility in association with migration distances and migrants' ages. The results suggest that as soon as Korea headed down the road to modernity, individual movements followed mixed migration patterns. Prior to modernization, migration in Korea exhibited both a stability-oriented pattern and a life-at-stake-oriented pattern. These findings confirm the context-specific diversity of migration processes across different societies and historical periods (Son & Lee, 2013). Migration patterns of people who lived in Seoul in the early 20th century were examined through age-specific migration rates and migration life tables using household registers (Kye & Park, 2013).

3.2 SOCIAL MOBILITY

Given the characteristics of the stratification system of the Joseon dynasty, social mobility is a critical issue in Korean history. The social stratification system in Korea has been called 'ambiguous' (Miyajima, 2003), because it had features of both the Japanese and Chinese models. Social status was not legally inheritable and depended on the reputation of the family. Social status played an impressive role in various ways. For example, social status was a more important criterion for being listed in a certain genealogy than lineage or birth order (Lee, 2010b). Therefore, elite family members had to make all kind of efforts to maintain their social status (Wagner, 1974b). Several studies examined the process and structure of reproduction of the elite family in pre-modern Korea. Using genealogical data, a study analyzed the hereditary tendency of bureaucratic reproduction in the 13th and 15th centuries by measuring the influence of fathers and grandfathers on the acquisition of official positions by individuals. According to this analysis, fathers had a strong influence on the acquisition of government posts, but the influence of grandfathers was low except for high-ranking government posts (Lee, 2013). In addition, there is a study that analyzed the intergenerational status mobility using family registers produced from the late 19th century to the early 20th century. Long-term mobility trends were identified based on the status of great-grandfathers, showing that absolute and relative mobility increased significantly in the late 19th century (Kye & Park, 2019).

Several researchers introduced methodology and theory enabling more advanced social mobility research. By linking genealogies to household registers, lineages can be linked to individual and household data on residence and social status (Kwon, 2014). This work challenges conventional views of intergenerational mobility of social status. Lee and Park (2018) constructed a prospective genealogical database containing all the records of public offices and family reproduction data over five generations of two elite family lineages in pre-modern Korea. They argue that the confluence of an ambiguous stratification system with a limited

number of high-ranking offices generated a trade-off for parents between the quantity and quality of positions attained by their offspring. The result of the trade-off was an unequal distribution of family resources aimed at reaching the lineage's collective goal, rather than maximizing the social rankings of individual children. Using a novel empirical strategy to consider the heterogeneous resource-allocation within elite families, this paper presents empirical evidence on associations between parents' and grandparents' social ranks and the quality of offices achieved by children of elite Korean families (Lee & Park, 2018).

4 FUTURE RESEARCH AT AIRG

As we have seen, Korean researchers have many achievements in social science history. Ongoing work at AIRG will extend historical and social science history through collaborations between historians and computer scientists (Lee, 2016b). A new field of research called 'digital history' is investigating the life courses of pre-modern Koreans, intergenerational- and multigenerational-effects of social mobility, the structure and function of the Korean family, the power mechanisms of the elite families, and other topics.

REFERENCES

- Cha, M. (2009). Jo-seon-hu-gi-ui chul-san-lyeog, sa-mang-lyeog mich in-gu-jeung-ga: Ne jog-bo-e na-ta-nan 1700–1899 nyeon-gan saeng-mol gi-log-eul i-yong-han yeon-gu [Fertility, mortality, and population growth in 18th and 19th century Korea: Evidence from genealogies]. *Korea Journal of Population Studies*, 32(1), 113–137. Retrieved from <http://www.koreascience.or.kr/journal/GOGHBY/v32n1.page>
- Choi, S., Choi, J., Paek, S. M., Yeh, H. J., & Lee, S. (2021). Jo-son-cho-gi gwan-ryo-us gwan-cheong-idong-eul tong-hai-bon ju-yo tong-chi-gi-gu-ui wi-sang: HAVNet ja-ryo-reul jung-sim-eu-ro [A study on the status of major government organizations based on the official movement of bureaucrats in the early Joseon dynasty: Focused on HAVNet Data]. *Sarim*, 75, 123–145. Available from <https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artid=ART002682892>
- Han, S. (2020). The historical background of the popularity of genealogies in Korea. *Journal of Family History*, 45(4), 498–516. doi: 10.1177/0363199020928364
- Hong, E., Lee, S., & Yoo, J. (2021). Strengthening the inner circle: The marriage network of elite families in Joseon Korea. *History of the Family*, 26, 313–335. doi: 10.1080/1081602X.2020.1869056
- Household Registers Research Team. (2003). *Danseong ho-jeok-dai-jang yeon-gu* [A study on Danseong Household Registers]. Seoul: Sungkyunkwan University Press.
- Jung, J. (2007). Yeog-sa-in-gu-hag ja-lyo-lo-seo-ui ho-jeog-dae-jang i-yong-eul wi-han gi-cho yeon-gu-tk dae-gu-bu-ho-jeog-dae-jang-gwa chon-lag-mun-seo-ui bi-gyo geom-to [A basic study on the use of the family register as a material for historical demography: Comparative review between 'Daegu family register' and village documents]. *Daedong Munhwa Yeon'gu*, 59, 363–401. doi: 10.18219/ddmh..59.200709.363
- Kim, K. (2005). Eighteenth-century Korean marriage customs: The Tansung census registers. *Continuity and Change*, 20(2), 193–209. doi: 10.1017/S0268416005005527
- Kim, K. (2009). Differing patterns of marriage between a city and villages in 18th century Korea: The case of Taegu area. *The History of the Family*, 14(1), 69–87. doi: 10.1016/j.hisfam.2008.12.002
- Kim, K., & Park, H. (2009). Landholding and fertility in Korea: 1914–1925. *Journal of Family History*, 34(3), 275–291. doi: 10.1177/0363199009337998
- Kim, K., & Park, H. (2010). Family succession through adoption in the Chosun Dynasty. *The History of the Family*, 15(4), 443–452. doi: 10.1016/j.hisfam.2010.09.002
- Kim, K., Park, H., & Jo, H. (2013). Tracking individuals and households: Longitudinal features of Danseong household register data. *The History of the Family*, 18(4), 378–397. doi: 10.1080/1081602X.2013.801357
- Kwon, K. (2014). Dal-seong seo-ssi-leul tong-hae bon jo-seon-hu-gi sin-bun-byeon-hwa-ui jang-gi chu-se-wa geu ui-mi: Telg-dal-seong-seo-ssi-jog-bo-telm-wa-telg-dae-gu-bu-ho-jeog-dae-jang-telm-eul jung-sim-eu-lo [A study on the long-term trend and significance of status changes in the second half of Joseon through the Dalseong Seo family: With a focus on genealogy of Dalseong Seo family and family register of Daegu]. *Hanguk Munhwa*, 67, 61–79. doi: 10.22943/han.2014..67.003

- Kwon, N. (2006). Jo-seon-hu-gi dong-seong-chon-lag gu-seong-won-ui tong-hon yang-sang: Dan-seong-hyeon sin-deung-myeon an-dong-gwon-ssi sa-lye [Marriage aspect of members of the single-lineage village in the late Joseon dynasty: Andong Kwons' case of Sundeung Myeon, Danseong Hyeon]. *The Journal of Korean History*, 132, 109–135. Retrieved from http://uci.kci.go.kr/resolution/result.do?res_cd=G704-000361.2006.132.005&res_svc_cd=
- Kwon, T., & Shin Y. (1977). Jo-seon-wang-jo-si-dae in-gu-chu-jeong-e gwan-han il-si-lon [On population estimates of the Yi dynasty, 1392–1910]. *Dong-a Munhwa*, 14, 289–330.
- Kye, B., & Park, H. (2013). Age patterns of migration among Korean adults in earth 20th-century Seoul. *The History of the Family*, 18(4), 398–412. doi: [10.1080/1081602X.2013.824910](https://doi.org/10.1080/1081602X.2013.824910)
- Kye, B., & Park, H. (2016). Marital fertility during the Korean demographic transition: Child survival and birth spacing. *The History of the Family*, 21(4), 483–501. doi: [10.1080/1081602X.2016.1183140](https://doi.org/10.1080/1081602X.2016.1183140)
- Kye, B., & Park, H. (2019). Intergenerational status mobility in nineteenth-century Korea: Evidence of Seoul household registers from 1897 to 1906. *Research in Social Stratification and Mobility*, 60, 52–65. doi: [10.1016/j.rssm.2019.03.001](https://doi.org/10.1016/j.rssm.2019.03.001)
- Lee, J. Z., & Campbell, C. D. (2016). *China Multi-Generational Panel Dataset, Liaoning (CMGPD-LN), 1749–1909 (ICPSR27063-v10)*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research. doi: [10.3886/ICPSR27063.v10](https://doi.org/10.3886/ICPSR27063.v10)
- Lee, S. (2010a). Go-lyeo-si-dae gwi-jog-cheung-ui sa-mang-lyul-gwa gi-dae-yeo-myeong-ui chu-se; Bi-gyo-sa-jeog gwan-jeom-eul jung-sim-eu-lo [The trend of mortality and life expectancy of the nobility in Koryŏ from comparative perspective]. *History & the Boundaries*, 47, 129–152. Retrieved from http://uci.kci.go.kr/resolution/result.do?res_cd=G704-001396.2010..74.007&res_svc_cd=
- Lee, S. (2010b). The impacts of birth order and social status on the genealogy register in thirteenth- to fifteenth-century Korea. *Journal of Family History*, 35(2), 115–127. doi: [10.1177/0363199009357158](https://doi.org/10.1177/0363199009357158)
- Lee, S. (2013). 'An-dong-gwon-ssi-seong-hwa-bo'e na-ta-nan 13–15se-gi gwan-lyo jae-saeng-san-gwa hyeol-yeon-gwan-gye [The impact of family background on bureaucratic reproduction in the thirteenth-to-fifteenth century Korea: A case study on the Andong Kwon-ssi Sunghwabo]. *Daedong Munhwa Yeon'gu*, 81, 41–67. doi: [10.18219/ddmh.81.201303.41](https://doi.org/10.18219/ddmh.81.201303.41)
- Lee, S. (2016a). Conditions and potentials of Korean history research based on 'big data' analysis: The beginning of 'digital history'. *The Korean Journal of Applied Statistics*, 29(6), 1007–1023. doi: [10.5351/KJAS.2016.29.6.1007](https://doi.org/10.5351/KJAS.2016.29.6.1007)
- Lee, S. (2016b). Towards a sustainable future for historical demography. In K. Matthijs, S. Hin, J. Kok & H. Matsuo, *The future of historical demography: Upside down and inside out* (pp. 245–248). Leuven/Den Haag: Acco. Retrieved from <https://soc.kuleuven.be/ceso/fapos/publications/the-future-of-historical-demography-upside-down-and-inside-out>
- Lee, S., & Lee, W. (2017). Strategizing marriage: A genealogical analysis of Korean marriage networks. *The Journal of Interdisciplinary History*, 48(1), 1–19. doi: [10.1162/JINH_a_01086](https://doi.org/10.1162/JINH_a_01086)
- Lee, S., & Park, J. (2018). Quality over quantity: A lineage-survival strategy of elite families in premodern Korea. *Social Science History*, 43(1), 31–61. doi: [10.1017/ssh.2018.38](https://doi.org/10.1017/ssh.2018.38)
- Lee, S., & Yoo, J. (2018). The unexpected effect of social status on reproduction: A case study in Joseon Korea from the nineteenth to the twentieth centuries. *The History of the Family*, 23(1), 109–134. doi: [10.1080/1081602X.2017.1338972](https://doi.org/10.1080/1081602X.2017.1338972)
- Miyajima, H. (2003). Jo-seon-si-dae-ui sin-bun, sin-bun-je gae-nyeom-e dae-ha-yeo [A study on the concepts of a person's status and the status system in the Joseon period]. *Daedong Munhwa Yeon'gu*, 42, 289–308. Retrieved from <https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artid=ART000887034>
- Miyajima, H. (2004). The present situation and the subject of Korean population history. *Sungkyun Journal of East Asian Studies*, 4(2), 1–9. Retrieved from <https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artid=ART00205284>
- Park, H., Han, S., & Kye, B. (2018). Changes in child mortality in Korea during the mid-twentieth century: Gender, birth order and sibling composition. *The History of the Family*, 23(4), 594–622. doi: [10.1080/1081602X.2018.1485114](https://doi.org/10.1080/1081602X.2018.1485114)
- Park, H., & Lee, S. (2008). A survey of data sources for studies of family and population in Korean history. *The History of the Family*, 13(3), 258–267. doi: [10.1016/j.hisfam.2008.05.005](https://doi.org/10.1016/j.hisfam.2008.05.005)
- Rhee, Y. (2004). A comparative historical study of the census registers of early Choson Korea and Ming China. *International Journal of Asian Studies*, 2(1), 25–55. doi: [10.1017/S1479591405000021](https://doi.org/10.1017/S1479591405000021)
- Son, B. (2007). *The household register 1606–1923: A cultural history of Joseon by means of recording population*. Seoul: Humanist Press.

- Son, B. (2010). The effects of man's remarriage and adoption on family succession in the 17th to the 19th century rural Korea: Based on the Andong Kwon clan genealogy. *Sungkyun Journal of East Asian Studies*, 10(1), 9–31. doi: [10.21866/esjeas.2010.10.1.002](https://doi.org/10.21866/esjeas.2010.10.1.002)
- Son, B. (2016). San ja-wa jug-eun ja-ui gi-jae-th-ho-jeog-gwa jog-bo-e dae-han yeog-sa-in-gu-hag-ui gwan-jeom-th [The records of the living and the dead: The viewpoint of the historical demography to household registers and genealogies]. *The Choson Dynasty History Association*, 79, 39–71. doi: [10.21568/CDHA.2016.12.79.39](https://doi.org/10.21568/CDHA.2016.12.79.39)
- Son, B., & Lee, S. (2010). The effect of social status on women's age at first childbirth in the late seventeenth- to early eighteenth-century Korea. *The History of the Family*, 15(4), 430–442. doi: [10.1016/j.hisfam.2010.09.001](https://doi.org/10.1016/j.hisfam.2010.09.001)
- Son, B., & Lee, S. (2013). Rural migration in Korea: A transition to the modern era. *The History of the Family*, 18(4), 422–433. doi: [10.1080/1081602X.2013.824909](https://doi.org/10.1080/1081602X.2013.824909)
- Wagner, E. W. (1974a). Social stratification in seventeenth-century Korea: Some observations from a 1663 Seoul census register. *Occasional Papers on Korea*, 1, 36–54. Retrieved from <http://www.jstor.org/stable/41490119>
- Wagner, E. W. (1974b). The ladder of success in Yi dynasty Korea. *Occasional Papers on Korea*, 1, 1–8. Retrieved from <https://www.jstor.org/stable/41490117>

HISTORICAL LIFE COURSE STUDIES
VOLUME 11 (2021), published 05-11-2021

The South African Families Database

Jeanne Cilliers

Lund University

ABSTRACT

Very little is known about what family life looked like for settlers in colonial South Africa during the 18th or 19th century, nor how events over these centuries might have affected demographic change. The primary reason for this lacuna is a shortage of adequate data. Historians and genealogists have, over the last century, worked to combine the rich administrative records that are available in the Cape Archives in Cape Town and beyond, into a single genealogical volume of all settlers living in the 18th, 19th and early 20th century. Until recently, this valuable resource was not in a format that would enable its use for the type of event-history analyses that have come to dominate the field of contemporary historical demography. This is now changing with the introduction of the South African Families database (SAF). SAF is one of very few databases known to document a full population of immigrants and their families over several generations. This article provides a brief background to, and technical overview of, the construction of the SAF. It discusses both the merits and limitations of its use in longitudinal demographic studies and offers a look into the types of studies it can enable.

Keywords: Historical demography, Genealogies, Longitudinal data, Life courses, Intermediate Data Structure, South Africa

DOI article: <https://doi.org/10.51964/hlcs11095>

© 2021, Cilliers

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

Assembling archival materials and historical registries to reconstruct family lineages of the European settlers to South Africa from the 17th to 20th centuries allows for an investigation into long-term economic and demographic trends across more than just two or three generations. Questions relating to the inter-generational transmission of socioeconomic status or about demographic processes such as fertility, migration, and marriage, that have previously gone unanswered, re-emerge.

Thanks to the wealth of documents kept by the Dutch East India Company (VOC) and the British colonial government when they ruled South Africa, much is already known about the establishment of the South African colonial society (Fourie, 2014). Less is known about what family life looked like for settlers in the 18th and 19th century nor how events over this period might have affected the way in which decisions around household formation were made. This is exacerbated to some extent by the fact that South Africa does not have a research hub for historical demography to encourage researchers to collect and transcribe data from the archives. As a result, South African historical demography remains in its infancy.

The South African Families Database (hereafter SAF) is a genealogical registry of settler families. It is one of very few in the world that is known to document a full population of immigrants and their families over several generations spanning nearly three centuries. The registers were painstakingly compiled by historians and genealogists using baptism and marriage registers, death notices, and individual family genealogies. The time-intensive nature of manual data transcription and a lack of computing power has meant that up until fairly recently, researchers opted to draw only small samples from these data, and as a result they had never been used in their entirety. Over the last decade these records have been turned into a functional database. The SAF database now includes information on all families known to have settled in South Africa and their descendants, complete until 1910, containing over half a million individuals.

This article provides a brief background to and technical overview of the construction of the South African Families database. It discusses both the strengths and limitations of its use in longitudinal demographic studies and offers a look into research currently being undertaken with these data at their core.

2 WHO WERE THE CAPE SETTLERS?

The Dutch, while not the first Europeans to ever traverse the southern parts of Africa — the Portuguese having done so a century prior — were the first to settle at the Cape of Good Hope, landing in 1652. In that year three ships of the VOC, under the Commander Jan van Riebeeck, arrived in Table Bay with the first company men. The VOC, with its base in Batavia, was a powerful monopolistic chartered company and the Cape was to serve the Company's ships as a rest stop on their passage to India. Of course, the Cape was not previously uninhabited. VOC company men settled on lands wrested from the indigenous Khoesan populations and their movements further inland were characterized by tension and violence between the groups.

Notions of family life amongst early European settlers at the Cape likely derived from the diverse cultural and religious practices of VOC employees' homelands. The end of Thirty Years War in 1648 saw European soldiers and refugees widely dispersed across the continent. Immigrants from Germany, Scandinavia, and Switzerland journeyed to Holland in the hope of finding employment and were amongst those who would make the six-month journey to settle the southern tip of Africa. Beyond this, the company filled its ranks with farm labourers, artisans, and unskilled workers from both rural and urban areas who spoke variations of French, Dutch, and German.

A consequential event of the 17th century at the Cape, was the arrival of about 170 French Huguenots in 1688 and 1689, by which time the free settler population had reached about six hundred. Cultural adaptation took place rapidly since new identities had to be shaped in a settler environment. De Kiewiet (1941, p. 6) described the arrival of the Huguenots as giving the Cape "more truly than before the contours and substance of a colony". He notes that although the Huguenots differed from the Dutch

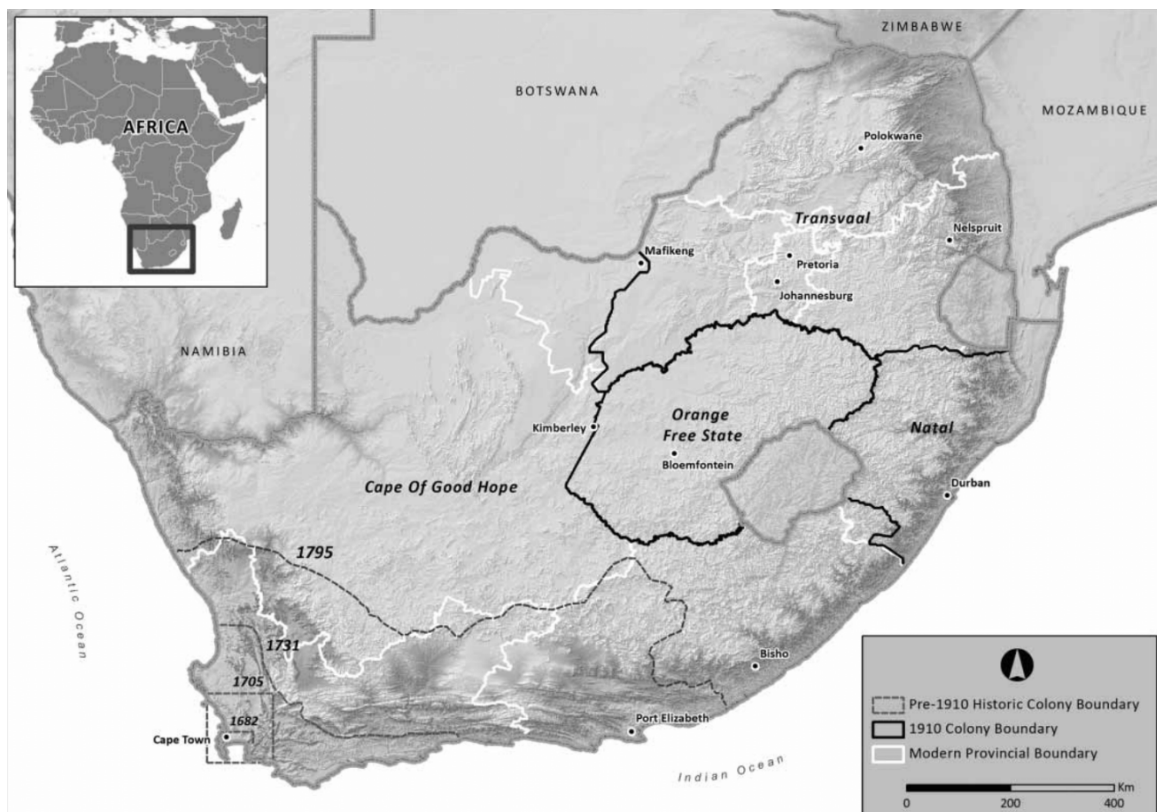
settlers in language, they were united by equal devoutness and tradition and "in two generations or less the groups had grown together and become one" (de Kiewiet, 1941, p. 6).

By the beginning of the 18th century free settlers had increased in number and influence and become more and more independent of the authority of Company officials. With the exception of the smallpox epidemics of 1713 and 1755, which resulted in slight declines in the population growth rate, the 18th century experienced a gross population growth rate of around 2.6% per annum (van Duin & Ross, 1987, p. 12). A steady flow of immigration of European settlers would continue so that by the end of the VOC's governance in 1795 the Colony was home to nearly 15,000 settlers (van Duin & Ross, 1987).

The British annexation of the Cape in 1795, and again in 1806 after a brief interlude of Batavian rule (1803–1806), brought immigrants from Britain to the Colony. Most notably some 4,000 settlers arrived in the Eastern Cape in 1820, as beneficiaries of a major scheme of assisted migration. The result of the arrival of an overtly British pressure group; the abolition of slavery; economic motives relating to insecure tenure of land and the abundance of fertile land beyond the frontier; and inadequate protection from native depredations were among the reasons cited for the mass exodus, beginning in 1835 of settlers further into the interior, known as the *Great Trek* (Neumark, 1957, p. 20).

These newly settled regions later formed the two independent Boer republics of the Orange Free State (1848) and the Transvaal (1852) and the colony of Natal (1843), which, together with the Cape Colony, became the four provinces of the Union of South Africa in 1910 (see Figure 1). The discovery of diamonds (1866) and gold (1886) in the two Boer republics boosted the population and income of settler South Africa. Migration to the diamond and gold fields increased rapidly, both from within the region and from outside its borders. Kimberley in the Orange Free State was the hub of the diamond industry, but its wealth was minor in comparison to the immense wealth generated by the discovery of gold in the Witwatersrand region of the Transvaal.

Figure 1 Map showing the settler expansion from the south-western Cape



Note: The expansion until 1795 is shown by way of the grey dashed lines; the four provinces that later constituted the Union of South Africa in 1910 by way of the black lines; and the modern-day provincial boundaries of South Africa, by way of white lines.

Source: Cilliers & Fourie, 2012.

While much is known about the political events of the pre-Union period, less is known about changes in living standards. The 17th- and 18th-century Cape Colony is generally considered to have been poor, almost entirely dependent on agriculture, although pockets of wealth could be found close to the market in Cape Town (Guelke & Shell, 1983). Recent scholarship has raised doubts about this view of the Cape Colony: Fourie (2013) uses probate inventories to show that 18th-century Cape settlers owned, on average, greater quantities of luxury goods than many of their European counterparts. Fourie and van Zanden (2013) find that Cape settlers' per capita income was in line with the most prosperous countries of the time, Holland and England.

How living standards, societal inequalities, and economic developments might have factored into individuals' choices about when and whom to marry, or what the ideal family size might be in this context, has previously been based on qualitative research. Anecdotal evidence seems to suggest exceptionally large family sizes. For example, Penn (2014) describes a woman in 1727, in her early 30s already the mother of seven children, who would go on to bear 11 children in total. Ross (1975) tells of a woman dying at the age of 49 at the birth of her twelfth child, whose husband would incidentally go on to father another 12 children with his second wife. While these cases appear to be outliers, they serve to highlight the difficulties of drawing conclusions based on a limited number of observations. Fertility rates for the early Cape Colony come from a study by Guelke (1988) in which the average number of children per woman are calculated from a sample size of fewer than 300, for just two years, 1705 and 1730. Simkins and van Heyningen (1989) offer similar snap-shot crude birth rate calculations using aggregated census data from 1891 and 1904 respectively. These aggregate censuses, available roughly decennially from the second half of the 19th century, while sufficiently broad in scope, do not allow for the possibility to follow individual households or family lineages over time. Evidently this body of literature requires an update. The SAF database is a springboard for a new generation of research that can address this lacuna.

3 THE SOURCE MATERIAL

The lineages that form the basis of SAF were compiled from thousands of source documents. According to the Genealogical Institute of South Africa (GISA) these sources include but are not limited to, baptism and marriage records of the Dutch Reformed Church archives in Cape Town; marriage documents of the courts of Cape Town, Graaff-Reinet, Tulbagh, Colesberg, collected from a card index in the Cape Archives Depot; death notices in the estate files of Cape Town and Bloemfontein; registers of the Reverends Archbell and Lindley; voortrekker baptismal register in the Dutch Reformed Church archive in Cape Town; marriage registers of the magistrate of Potchefstroom; other notable genealogical publications including *Geslachtregister der Oude Kaapsche Familien* [Genealogies of Old Cape Families] (De Villiers, 1894); *Die Herkoms van die Afrikaner* [The Origins of the Afrikaner], 1657–1867 (Heese, 1971); *The Family Register of the South African Nation* (Malherbe, 1966); *Some Frontier Families* (Mitford-Baberton & White, 1968), and various individual families genealogical publications.

Varying degrees of measurement error may have been introduced during the process of data compilation and digitization. The first is the possibility of errors in the original source documents. Misspelling of names and misreporting of dates are likely, given the differential precision applied by the members of the clergy and colonial administration responsible for the maintenance of the respective records. Next, mistakes will have inevitably cropped up in the process of compiling the genealogies. Many of the source documents were copies of originals that had been lost, in some the writing was faded, indistinct or illegible, or had already been transposed a number of times. The degree to which genealogists made discretionary choices in such instances can never be fully known. These issues will be handled systematically in section 5.

The resulting volumes, *South African Genealogies* (2008) and *South African Families* (2012) represent over a century of effort by South African genealogists, many of whom devoted their careers to creating and expanding these registers. In doing so they have, perhaps unintentionally, provided a rich source for exploring South African settler demographic history. An excerpt from the Cilliers lineage (Figure 2) shows the format of a typical register. A short text biography of the progenitor is provided, often containing some details about his region of origin or journey to the Cape. In this example, we are told

that Josué Cellier was born in 1667 in Orleans, France, and arrived at the Cape, aboard a vessel named the "Reygersdaal" in 1700, together with his wife, Elisabeth Couvert whom he had married that same year. They settled on "Het Kruyspad" farm in the district of Brackenfell and later moved to "Orleans" in Daljosaphat. She would go on to marry Paul Roux in 1722 after Josue's death in 1721. Their (Josue and Elisabeth's) children are listed below.

Figure 2 Excerpt from 'South African Families' (2012)

CELLIERS / CILLIERS / CILLIE

Josué Cellier * Orleans, Frankryk c. 1667

a. aan Kaap 1700 aan boord Reygersdaal met sy vrou, vestig aanvanklik te "Het Kruyspad", dist Brackenfell en Later "Orleans", Daljosaphat. Volgens Boucher is Josue Celliers moontlik die seun v Josue Celliers en sy vrou Judith Rouilly. Hierdie egpaar het 'n seun Nicolaas in die kerk te Bazoches-en-Dunois laat doop. † "De Orleans", dist Drakenstein Okt. 1721 x Frankryk c. 1700, Elisabeth COUVERT * Orleans, Frankryk c. 1676 † c. 1743 (sy xx c. 1722 Paul Roux † Drakenstein 7.2.1723)

b1 Josué ≈ Drakenstein 2.1.1701 † dist. Drakenstein 19.4.1770, ongetroud

b2 Jan ≈ c. 1702 † c. 1755, burger v Drakenstein x Paarl 5.12.1728 Anna MARAIS * ≈ c. 1707 (wed. v. Gabriel Rossouw) † dist. Drakenstein 11.1.1765 d.v. Charles Marais en Anna de Ruelle

c1 Jan ≈ Paarl 9.10.1729 † dist. Drakenstein 6.6.1766 x Tulbagh 8.10.1751 Susanna MALHERBE ≈ Drakenstein 15.2.1733 † c.1754 d.v. Pierre Malherbe en Elisabeth Cellier xx Paarl 11.7.1756 Sara Margaretha ROSSOUW ≈ Drakenstein 5.8.1736 † Drakenstein 18.7.1821 d.v. Daniel Rossouw en Sara Hanekom

d1 Johannes ≈ Paarl 30.7.1752 † dist. Drakenstein 5.6.1816 x Paarl 12.10.1783 Anna Maria NAUDE ≈ Paarl 12.10.1760 † dist. Drakenstein 22.7.1809 d.v. Jacob Naude en Susanna du Toit

e1 Johannes Francois ≈ Paarl 19.9.1784 † dist. Paarl 10.9.1843 x Paarl 13.6.1806 Anna Magdalena ROSSOUW * 2.3.1788 ≈ Paarl 9.3.1788 † dist. Drakenstein 7.7.1822 d.v. Pieter Rossouw en Anna Cilliers xx Cradock 5.12.1824 Maria Magdalena BREED * 7.2.1807 ≈ Graaff-Reinet 26.4.1807 d.v. Johannes Augustus Breed en Johanna Venter

f1 Anna Magdalena * 12.7.1807 ≈ Paarl 9.8.1807 † Prince Alfred Hamlet 6.6.1873 x Paarl 5.4.1829 (Johannes) Cornelis Jeremias GOOSEN ≈ 10.9.1809 † 6.10.1892 s.v. Gideon Jacobus Goosen en Hester Catharina Malan

f2 Johannes Francois * 13.3.1810 ≈ Paarl 8.4.1810 † Paarl 31.10.1879 x Paarl 8.9.1835 Maria Johanna DU TOIT * c. 1815 † Paarl 9.6.1874 d.v. Daniel du Toit en Maria Elizabeth Marais

4 DATA CAPTURING AND CODING

At the outset, transforming these registers into a functional format fit for analysis proved an enormous task, which spanned the better part of 2011. The first step in the data capturing process was to create a custom-designed data-transposing software that was able to convert what was essentially long text strings demarcated with proprietary symbology into a format compatible with conventional statistical software packages which also captured only the relevant information. This was a cumbersome process as the programme, while innovative, was not able to distinguish between successive families and meant that data had to be read in on a family by family basis. Resulting from typesetting inconsistencies in the SAF volumes, many records still required substantial post-transcription cleaning. Some family lineages, the Cilliers for example, were compiled by the Genealogical Institute of South Africa (GISA) in Afrikaans while others were in English. For consistency, all output was translated to English.

The first resulting dataset captured only the following individual-level information: names, surnames, birth, baptism, marriage and death dates and position in the genealogical tree. Soon after this initial phase of transcription, however, GISA undertook to revise and republish the registers, with the aim of correcting errors where possible, and extending the series to contain complete family registers of all

settler families up to 1930. A new edition of the genealogical registers was published by the GISA in 2014 and contained complete family registers of all settler families from 1652 to approximately 1830 as well as those of new progenitors of settler families up to 1867 for families with surnames starting with the letters A–Z, and up to 1930 for families with surnames starting with the letters A–K. In 2016, GISA completed their revisions of L surname families before bequeathing the ownership and copyright of the series to the Genealogical Society of South Africa (GSSA). Although several registers M–Z have subsequently been updated, and continue to be revised, the latest release of the SAF database only includes up to the L revision.

To transcribe these new versions of the registers a more sophisticated data transcription programme was designed to extract more information. This process was completed in 2013 and a new dataset containing both the original set of variables, as well as new information on occupation (where available), locations of vital events, and spousal information including birth, baptism and death dates and places as well as maiden names (where applicable) and parents' names. The inclusion of the new information was limited to surnames starting with A–K information but provided a significant sample size increase. The inclusion of the revised and expanded A–K data into the original dataset was permitted since having a surname A–K was not found to make an individual systematically different from those with surname starting with L–Z on observable characteristics, including age at first marriage and net fertility. Moreover, no systematic differences between the two versions of the data, other than the increased sample size, indicate that any errors that might remain in the data can be safely attributed to the underlying data, rather than the transcription process. The SAF database is freely available for academic use. An anonymized version of the database will be made publicly available while use of the full version (including personal identifiers) can be made available upon request.

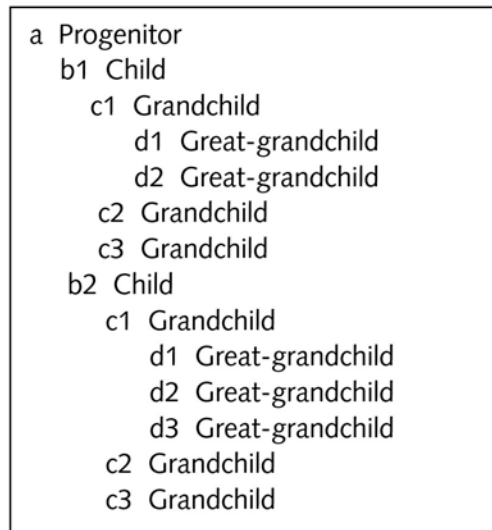
As will become apparent, additional variables were critical to enable the broader usability of the dataset for the purposes of longitudinal or event-history analysis. This is because the original structure of the genealogies is patrilineal. That is, children appear under their father's lineage and are not directly linked to their mothers. In the first example in Figure 2, this would not be problematic because a list of all the offspring of Josué and Elisabeth is given. If, however, Josué had remarried and continued to have children with a second wife, the listed offspring would have to be assigned to their respective mothers. To do so, information about death and/or marriage dates of the spouses is needed. However, this is not always available which limits research into for example female fertility.

In addition to the information captured directly from the source, a number of new variables were generated during this stage of transcription to facilitate the linking of individuals to both of their parents, and the tracing of familial relationships with relative ease over multiple generations. To generate unique individual identity codes, genealogical codes were concatenated to surnames to indicate the relative position of individuals on their family tree. The genealogical codes follow the de Villiers-Pama numbering system. The de Villiers-Pama System is similar to the Henry Numbering System more commonly used in the United States, except that each digit (or group of two digits for numbers larger than 9) is preceded by a generation letter. The progenitor of a particular family, or the first ancestor of that family entering the country is assigned the letter "a". This designates him or her as the "a" generation. The "a" is followed by a number showing which child he/she was. "a3" would mean that the person was the third child in the "a" generation. The children of a3 will be the "b" generation. They will be numbered according to how they were born — the eldest or first born being b1; the second b2; b3 etc. Children descending from the "b" generation will be the "c" generation and so on. An illustration is provided in Figure 3.

On top of the unique individual ID codes assigned, individuals were also assigned a sibship ID, which equates to their individual ID with the last digit (or group of two digits for birth orders higher than 9) removed. The final two entries from the excerpt in Figure 2, Anna Magdalena and Johannes Francois, would therefore have individual IDs: CILLIERS_a1b2c1e1f1 and CILLIERS_a1b2c1e1f2, respectively and would share the sibling ID: CILLIERS_a1b2c1d1e1f. Their father is identifiable by removing the last character from their sibling ID: CILLIERS_a1b2c1d1e1.

Sex is never stated in the records nor is it immediately discernible from the de Villiers-Pama genealogical coding. A sex variable was generated for every individual post-transcription, through a semi-automated process whereby sex was attributed based on the likelihood that first and second names of individuals and their spouses matched a predetermined list of common South African names. All ambiguous cases were dealt with manually. Individuals for whom a sex was indeterminable constitute around 1% of the database.

Figure 3 Example of the de Villiers-Pama structure



Since women appear as wives in their husband's households but are not directly linked to their own children through the transcription process, a mother's ID variable was generated and assigned to each individual. In cases where a man was only married once in his lifetime (94.5% of the fathers in the sample), matching mothers to their children was a relatively straightforward process using the individual and sibling identification codes. In these cases individuals who share a sibling identifier all are assigned the same mother's ID (the ID of their father's only wife). Cases where men married more than once require more careful distinction of children belonging to the first wife from children belonging to the second, third, or in some rare cases, fourth wife. An algorithm using the previous wife's death date, subsequent marriage date, and the birth dates of all of the children, allows for the linking of children to the correct mother. In the event that there was more than one wife and a birth or death date was missing, a successful match cannot be made. As a result, 18% of non-progenitor individuals in the database have a missing mother's ID in the database.

With all familial relationships clearly established in the database, conversion to longitudinal format to allow for event-history analysis, was fairly straightforward. The only familial relationship that remains untraced is that of children to their mother's ancestors. This is because females appear as children in their father's genealogy i.e., under their maiden names, and then as wives in their husband's genealogy, i.e. under their married names. The possibility to make this linkage does exist using these maiden names, and record-linkage strategies to do so are currently being explored.

5 REPRESENTATIVENESS

Specific concerns related to data appropriateness or representativeness for a given research question are already covered extensively in the various publications which make use of the SAF database. These will be discussed briefly below but a few general points regarding data quality are worth making here.

Family lineages have long been used by demographers in their studies on past demographic behaviour. The common problems associated with the use of genealogical data in historical demography research are already well documented (Hollingsworth, 1969; Willigan & Lynch, 1982; Zhao, 2001) and they are obviously biased towards the fertile and the marriageable. By definition, a genealogy is the written record of a family descended from a common ancestor or ancestors, and as a result, most genealogies are the records of members of *surviving* patrilineages. These families would most likely have experienced favourable demographic conditions which resulted in their survival. The use of these genealogies may, therefore, not be representative of the history of the whole population in question (Zhao, 2001, p. 181). As Willigan and Lynch (1982, p. 112) argue: "Genealogies were often designed to emphasize not only the glorious aspects of a lineages past but also its durability through time. Consequently, members who contributed little to the group's duration were likely to be missing or underrepresented. This category might include individuals who did not reach maturity and those who

survived but had no children, or who had children who themselves died at a young age or failed to reproduce. This creates a bias towards long generations (late marriage, remarriage, late child-bearing, high fertility) and long life." In general, the greater the number of generations recorded, the smaller the impact of the selective bias, as long as the genealogy does not suffer severely from other types of under-registration. If the genealogy is shallow in generational depth or the members of the first few generations consist of a large part of the population being investigated, the selective biases are more likely to affect the outcome. Otherwise, their influences can be negligible. The SAF database benefits from great generational depth (see Table 1). However, as a result of small population sizes (the entire free burgher population consisting of less than 1000 individuals before 1700) and very small sample sizes for the period 1652–1699, using SAF to study the period prior to 1700 is not advised.

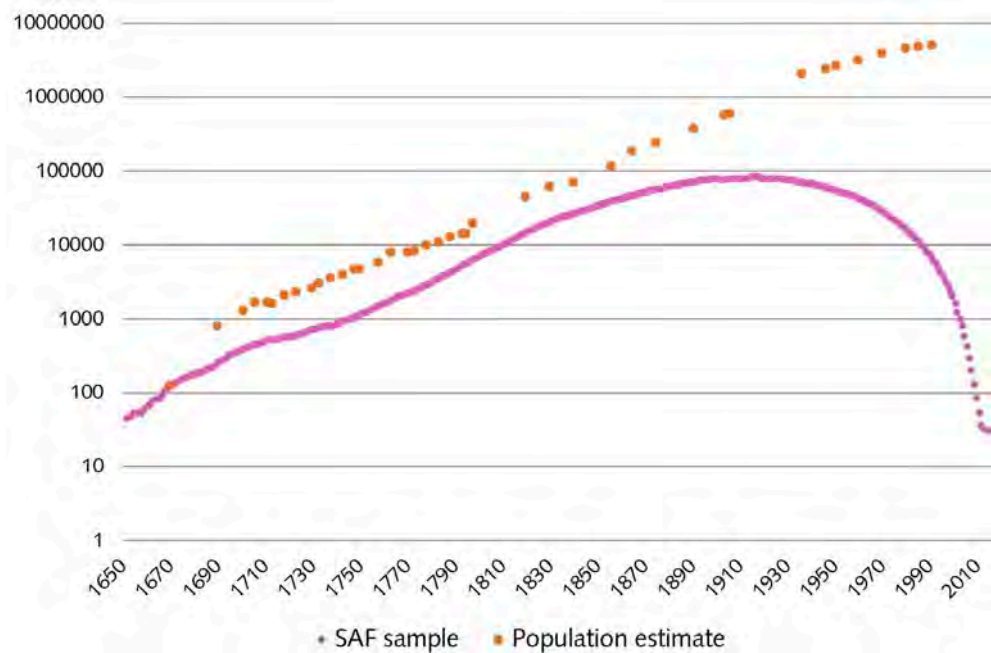
Table 1 *Distribution of individuals across generations, by birth cohort*

Generation	1650–1749	1750–1849	1850–1949	Total	% of sample
1	204	1,624	4,343	6,171	1.40
2	2,599	15,641	22,516	40,756	9.27
3	1,988	14,569	24,024	40,581	9.23
4	627	19,316	28,388	48,331	10.99
5	25	26,913	34,913	61,851	14.07
6	1	27,391	52,739	80,131	18.23
7	0	9,862	77,971	87,833	19.98
8	0	767	52,701	53,468	12.16
9	0	18	17,277	17,295	3.93
10	0	1	3,019	3,020	0.69
11	0	0	154	154	0.04
12	0	0	17	17	0.00
Total	5,444	116,102	318,062	439,608	100.00

It is also necessary to address the representativeness of the SAF data in terms of the size of the documented historical population. While GISA asserts that the registers are complete up until 1869 for all families and complete to 1930 for families with surnames starting with letters A–L, the registers also contain information on individuals up to the present. This information only exists, however, where families have taken it upon themselves to keep information on their family trees publicly up to date. This calls into question the representativeness of the registers after 1930, since it is unclear what kind of a bias this self-selection into the registers would introduce.

Moreover, as illustrated by Figure 4 which plots the sample size against the actual population over the whole period, the sample closely correlates with estimates of the total settler population for the 18th century and 19th century. By the early 20th century absolute SAF sample size slows considerably relative to the total settler population, and by roughly 1912, the sample size reaches a turning point and begins to decrease in size.

The year 1910 which marks the political unification of the two British colonies, the Cape Colony and Natal, and the two Boer republics, the Orange Free State and the South African Republic, seems an appropriate year up to which this sample could be used as a representative source of information on European settlers and their descendants in South Africa. A further limitation is that SAF do not follow individuals who emigrated from South Africa, nor is there any clear way of discerning outmigrants from those whose lineages ended for other, unrelated reasons. The year 1910 as a cut-off point is additionally useful, since it precludes users from possible violations of privacy regulations which protect the data for a period of one hundred years.

Figure 4 *Sample size versus population estimate*

Note: Log scale. Population size provided for years for which a population estimate is available.

Sources: *Census of the colony of the Cape of Good Hope, 1856, 1865, 1891, 1904, 1910; Elphick & Giliomee, 1989; Ross, 1975; Sadie, 2000; and own calculations.*

Beyond mirroring the general trend in population growth, Cilliers and Mariotti (2019) provide further comparisons of the age, sex, and regional distributions of SAF to census data, where possible, confirming that the database does not suffer from systematic compositional bias.

Of additional concern is partial or incomplete data on individuals. While the size and scope of the SAF data are its greatest advantage, it must be noted that not all entries contain complete information. Of the full dataset, which contains 671,385 observations, many entries are empty save for a name and surname. Close to two thirds of these entries contain a birth or a baptism date, while only one quarter contains a death date, and less than one fifth contains a marriage year. These statistics can be found in Table 2.

When individuals whose data are partial or incomplete are removed from the study in question, the sample size is substantially reduced. If we consider the SAF sample for which there are complete birth and death dates, it effectively captures approximately 30% of the total estimated population over time, reducing from around 1865 to about 10% around 1910 (see Figure 5). In addition, if there is a systematic relationship between the demographic event under investigation and the likelihood that information is incomplete, this will introduce additional bias to the study.

Table 2 *Frequency of observations in the dataset for selected variables*

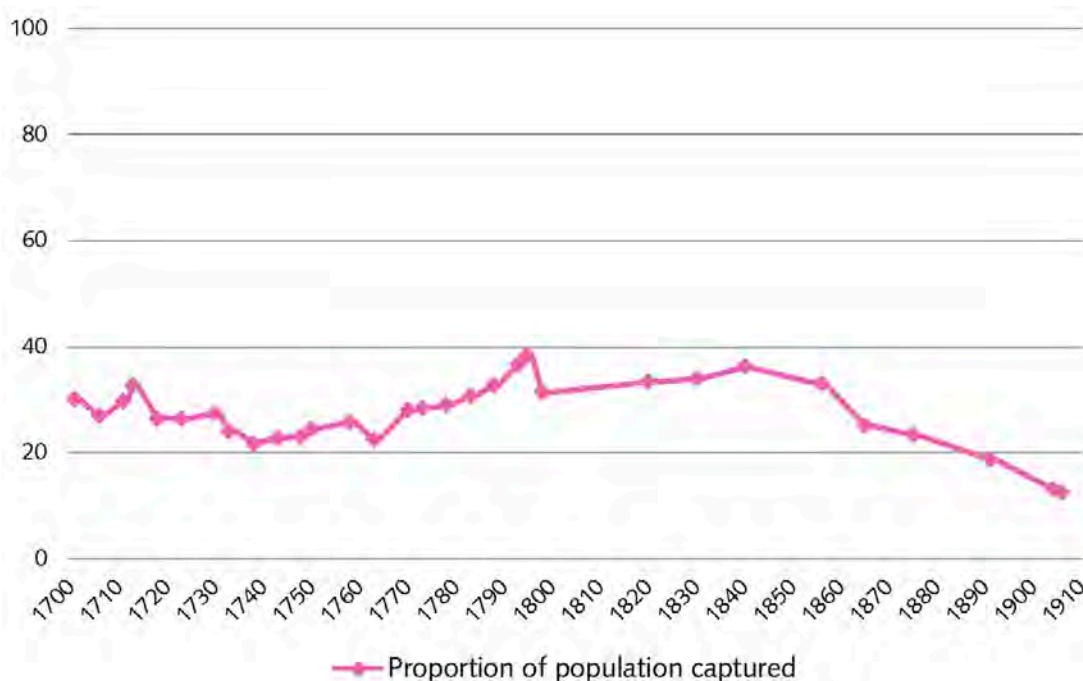
Variable name	Females	Males	Sex unknown	Total
Individuals	305,260	360,936	5,189	671,385
Individuals with known fathers	300,625	333,214	4,251	638,090
Individuals with known mothers	238,757	276,724	2,399	517,880
Individuals with known year of birth/baptism	219,089	255,902	2,029	477,020
Individuals with known year of first marriage	67,602	94,115	119	161,836
Individuals with known year of death	40,978	98,894	459	140,331
Individuals with all data known	13,771	32,977	12	46,760

6 SOCIO-ECONOMIC STATUS

The only measure of socio-economic status provided in SAF are occupations. Occupations in the database have been coded according to the Historical International Standard Classification of Occupations, hereafter HISCO (van Leeuwen, Maas, & Miles, 2002), and then classified according to the Historical International Social Class Scheme, hereafter HISCLASS (van Leeuwen & Maas, 2011). Bias arising from the incomplete or inconsistent reporting of occupations is of particular concern. However, comparisons with available census data reveal that the reporting of occupations do not appear to be systematically related to the relative ranking of certain occupations in society.

Comparing the white working age male population from Cape Colony censuses to estimates for equivalent time-periods from SAF in table 3, shows that although discrepancies exist between the SAF and the true occupational structure of the population, the general levels and trends are correlated. Still, occupations are typically not reported for women, and roughly only 10% of men in SAF have one or more occupations listed chronologically (not associated with a specific date or individual's age), providing a less than ideal measure of socio-economic status.

Figure 5 SAF sample with complete birth and death dates as a proportion of the total settler population, 1700–1908



Sources: *Census of the colony of the Cape of Good Hope, 1856, 1865, 1891, 1904, 1910*; Elphick & Giliomee, 1989; Ross, 1975; Sadie, 2000; and own calculations.

Table 3 Share of the European/white working age male population with specified occupations, from available Cape of Good Hope censuses, compared to SAF by skill group

Skill group	1850 SAF	1865 Census	1900 SAF	1911 Census
White collar	20.5	29.7	33.4	29.3
Farmer	64.4	55.3	49.8	47.8
Skilled/semi-skilled	8.4	7.5	13.9	19.0
Unskilled	6.8	7.5	3.0	3.8
N	1,602	48,485	5,327	493,562

7 LINKING TO EXTERNAL SOURCES

7.1 MANUAL RECORD LINKAGE: PROBATE INVENTORIES

Given the limitations of SAF in terms of refined socio-economic variables, additional sources can help to complete the database (see Figure 6 for the time-coverage of supplementary datasets). These required the development of suitable record linkage strategies. The database has already been manually supplemented with information from probate inventories compiled by the Master of the Orphan Chambers (MOOC). The Orphan Chamber was set up in 1673 and operated until 1834 and the inventories of the Orphan Chamber (MOOC 8-series) are an invaluable source for researchers interested in the lives of people at the early Cape. The inventories list all the possessions in a deceased estate, including livestock and slaves, and were a relatively complete and undisturbed reflection of households at the time of appraisal, which usually took place within days of death. In the rural districts, possessions were inventoried by neighbours, relatives or friends and sent to Cape Town. A clerk then copied the appraisal into a standard format, though the original details were retained (TANAP, 2010). The MOOC 8-series was manually linked to SAF based on individuals' unique first name(s) and surname strings and their birth and death dates (where available) resulting in the linkage of 2,117 of the 4,160 probate inventories, representing just over 50% (Fourie & Swanepoel, 2018).¹ This has proven to be a valuable addition to the database enabling the study of intergenerational transmission of wealth, crucially, for both males and females (Cilliers, Fourie, & Swanepoel, 2019).

Beyond their material wealth, Cape settlers held substantial shares of wealth in slaveholdings. The slave valuation and compensation records from 1834, the year slave abolition was enacted by the British Empire, can be found in the Cape Town Archives. They contain information on slaves (names, sex, age, place of birth, and value) and the slaveholders. Martins, Cilliers and Fourie (2020) manually linked slaveholder data to SAF for the Stellenbosch district. Linkage for the remaining districts of the colony is currently underway.

7.2 AUTOMATED RECORD LINKAGE: TAX CENSUSES

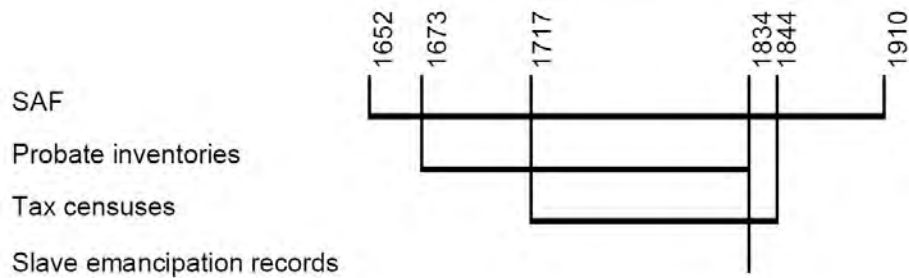
Further supplementary data comes from the *opgaafrollen*, annual tax censuses collected between 1663 and 1844, first by the Dutch East India administration and after 1795 by the British colonial administration, of all free households of the Colony. Household-level information includes name and surname of the head of the household and spouse, the number of children present in the household, the number of slaves and indigenous Khoisan employed, and several agricultural inputs and outputs, including cattle, sheep, horses, wheat sown, wheat reaped, vines, and wine produced. The series of *opgaafrollen* housed in the Cape Archives (1717–1844) are in the process of being transcribed under the umbrella of the Cape of Good Hope Panel project, a joint venture between Lund University and Stellenbosch University (Fourie & Green, 2018).

The *opgaafrollen* returns are, themselves, in the process of being linked across years to create an annual panel of household production. The data from one district served as the pilot study for the development of an automated probabilistic record linkage strategy that will soon be rolled out to the remaining districts (Rijpma, Cilliers, & Fourie, 2020). At the time of writing, individuals in SAF have been linked to household heads in the *opgaafrollen* for only the Graaff-Reinet district of the Cape of Good Hope Panel. Once complete, the Cape of Good Hope Panel will be the longest dataset of its kind in existence, spanning a period that stretches beyond any one lifetime. The inclusion of inputs such as household size and labor employed and outputs such as grain, wine and stock, allows for the testing of theories of economic growth and development, labor markets, industrial organization, political economy, migration and institutional economics, but also to develop and apply new econometric techniques.

The combination of the Cape of Good Hope Panel with SAF allows for a number of novel multigenerational studies. Firstly, it contains a heterogeneous group of individuals or households, whose behavior may be vastly different even within the same region. Secondly, following households over time allows for the study of reactions to changes in economic and social circumstances or (exogenous) institutional changes. With an intergenerational panel even more could be done: how these exogenous shocks affect families over multiple generations, and whether these processes are time-persistent and dependent on the initial conditions from which those families started out could be ascertained.

¹ Non-unique name and surname combinations make linkage impossible since the correct individual cannot be selected from a list of possible candidates. Still, a linkage rate of 50% is generally considered to be high for historical data.

Figure 6 Coverage of supplementary datasets



Note: The line for the tax censuses represents the potential for linkage (only one pilot district has, at the time of writing, been successfully linked: Graaff Reinet district, 1786–1834), while all the other lines represent completed linkage between sources.

8 LOCATIONS

Certain dimensions of SAF remain unexplored. Locations are just one. Just over one third of all reported birth/baptism dates are accompanied by location information (residence of parents for the births and place of registration for the baptisms). For the most part these data are not yet cleaned or geocoded and therefore not yet available for public use. Where previous studies have made use of the location information from SAF, it has been made possible by drawing sub-samples from the database, for which location information was processed and assigned a relevant broader district categorization in lieu of exact co-ordinates. Geocoding all location information available in SAF is one of the priority steps to be taken in the further development of the database.

9 SELECTED STUDIES USING SAF

Genealogical data are particularly useful for the study of individuals, families, or communities across multiple generations. They are additionally well-suited for cohort analysis (Hollingsworth, 1969) since individuals belonging to the same cohort will have typically experienced the same vital event, birth or marriage for example, during the same period. With these advantages in mind and since data constraints define the limitation of studies, it is useful to ask which types of questions these data are best-suited to answer. Given the relative completeness of birth recording and the capacity to link individuals across generations, the obvious topics are fertility and intergenerational mobility.

Cilliers and Mariotti (2019) provide a complete series of female fertility estimates in the Cape Colony from 1700 to 1909. While previous research used portions of these data (Cilliers & Fourie 2012; Gouws, 1987) this was the first paper to use the full SAF database to date the onset of the South African fertility transition, which was found to have begun in the late 1870s to women born in the 1850s. Cilliers and Mariotti (2021) take further advantage of the longitudinal nature of the database to revisit the discussion on family limitation through stopping and spacing behavior prior to and during the fertility transition. Using split population estimation (cure models), the study finds that physiology and fecundity were the main determinants of both stopping and spacing behaviour prior to the fertility transition. The paper does not find evidence of explicit parity-dependent control for either stopping or spacing although some evidence of variation in birth interval lengths driven by postponement is found. During the transition, an increase in both stopping behavior and variation in birth interval length driven by postponement is found, followed by an increase in spacing after the transition.

Exploiting the intergenerational character of the data, Piraino, Mullier, Cilliers and Fourie (2014) investigate the intergenerational transmission of longevity between parents and offspring and find a positive and significant association between parents' and offspring's life duration, as well as between siblings. While these correlations persist over time, the magnitude of the effect is relatively small.

The effect of grandparents' longevity on that of grandchildren is insignificant, but cousin correlations suggest that inequality in longevity might persist across more than two generations. It was suggested that family and environmental factors shared by cousins could explain these results.

10 INTERMEDIATE DATA STRUCTURE

The SAF userbase has, for the most part, been limited to a handful of scholars interested in revising the existing historiography of the Cape Colony. While this is by no means an undeserving objective, this article serves to demonstrate that the value added by SAF could extend far beyond this. Most striking is the dearth of comparative studies emerging from this database. To attract a broader userbase of international historical demographers, and to facilitate comparative demographic studies, a standardization of sorts was warranted. It was for this reason that the decision was made to transfer SAF into the Intermediate Data Structure (IDS) (Alter & Mandemakers, 2014).

While IDS is not the only way to store and extract data, it is favoured by a growing number of longitudinal historical databases, not least because of its simple format that can suit many different types of data but also because it solves many of the problems related to the time-dependent nature of most historical demographic data. The principles of IDS involves two layers of data: 1) data about individuals and relations between individuals, and 2) data about contexts and relations between individuals and contexts. This design yields the five principal IDS tables. The INDIVIDUAL table, containing individual attributes; the INDIV_INDIV table, containing individual relations; the CONTEXT table, containing attributes of a geographical space where individuals reside together; the CONTEXT_CONTEXT table, containing how contexts are related to one another; and finally, the INDIV_CONTEXT table, relating the two layers of data.

Following the step-by-step guide provided by Klancher Merchant and Alter (2017), ENTITY and RELATIONSHIP tables were created for SAF. These are the necessary files to enable use of the IDS Transposer — an online tool which automatically transforms prepared data into the IDS standard. This produced two of the five IDS tables mentioned above: INDIV and INDIV_INDIV, since location information in SAF is not yet ready for wider use. These two “pilot” tables are now available for public use and following further development of the database it is hoped that full use of the IDS standard is on the horizon.

11 CONCLUSION

By shedding new light on the demographic characteristics of European settlers in 18th-, 19th- and early 20th-century Cape Colony, a severely under-researched topic in South African economic history, we can begin to move beyond a mere restatement of the history, producing results which not only challenge the existing understanding of South African historiography, but which add to the international debate around the nature and causes of demographic transitions. The limitations of genealogical data in terms of national representativeness and under-enumeration bias cannot be overlooked, but the research possibilities which capitalise on its highly valuable longitudinal and individual-level properties are boundless.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my various co-authors, Johan Fourie, Erik Green, Martine Mariotti, Igor Martins, Sean Millier, Patrizio Piraino, Auke Rijpma, and Christie Swanepoel who have contributed their time, effort, and expertise towards making the SAF database a functional source for future research.

I am grateful to George Alter and Luciana Quaranta for their advice and guidance with IDS conversion.

REFERENCES

- Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies*, 1, 1–26. doi: [10.51964/hlcs9298](https://doi.org/10.51964/hlcs9298)
- Cape of Good Hope (South Africa). (1856). *Census of the colony of the Cape of Good Hope. 1856*. Cape Town: Solomon.
- Cape of Good Hope (South Africa). (1866). *Census of the colony of the Cape of Good Hope. 1865*. Cape Town: Solomon.
- Cape of Good Hope (South Africa). (1891). *Census of the colony of the Cape of Good Hope. 1891*. Cape Town: Solomon.
- Cape of Good Hope (South Africa). (1905). *Census of the colony of the Cape of Good Hope. 1904*. Cape Town: Government Printer.
- Cape of Good Hope (South Africa). (1911). *Census of the colony of the Cape of Good Hope. 1910*. Cape Town: Government Printer.
- Cilliers, J., & Fourie, J. (2012). New estimates of settler life span and other demographic trends in South Africa, 1652–1948. *Economic History of Developing Regions*, 27(2), 61–86. doi: [10.1080/20780389.2012.745663](https://doi.org/10.1080/20780389.2012.745663)
- Cilliers, J., & Fourie, J. (2014). Die huwelikspatrone van Europese setlaars aan die Kaap, 1652–1910. *New Contree*, 69, 45–70. Retrieved from <http://dspace.nwu.ac.za/handle/10394/10906>
- Cilliers, J., & Fourie, J. (2018). Occupational mobility during South Africa's industrial take-off. *South African Journal of Economics*, 86(1), 3–22. doi: [10.1111/saje.12177](https://doi.org/10.1111/saje.12177)
- Cilliers, J., Fourie, J., & Swanepoel, C. (2019). 'Unobtrusively into the ranks of colonial society': Intergenerational wealth mobility in the Cape Colony over the eighteenth century. *Economic History of Developing Regions*, 34(1), 48–71. doi: [10.1080/20780389.2019.1574565](https://doi.org/10.1080/20780389.2019.1574565)
- Cilliers, J., & Mariotti, M. (2019). The shaping of a settler fertility transition: Eighteenth- and nineteenth-century South African demographic history reconsidered. *European Review of Economic History*, 23(4), 421–445. doi: [10.1093/ereh/hey019](https://doi.org/10.1093/ereh/hey019)
- Cilliers, J., & Mariotti, M. (2021). Stop! Go! What can we learn about family planning from birth timing in settler South Africa, 1835–1950? *Demography*, 58(3), 901–925. doi: [10.1215/00703370-9164749](https://doi.org/10.1215/00703370-9164749)
- de Kiewiet, C. W. (1941). *A history of South Africa, social & economic*. Oxford: Clarendon Press; New York: Oxford University Press.
- De Villiers, C. C. (1893). *Geslacht-register der oude Kaapse familien*. Kaapstad: Van de Sandt de Villiers & Co.
- Elphick, R., & Giliomee, H. (Eds.). (1989). *The shaping of South African society, 1652–1840* (1st Wesleyan ed.). Middletown, CT: Wesleyan University Press.
- Fourie, J. (2013). The remarkable wealth of the Dutch Cape Colony: Measurements from eighteenth-century probate inventories. *The Economic History Review*, 66(2), 419–448. doi: [10.1111/j.1468-0289.2012.00662.x](https://doi.org/10.1111/j.1468-0289.2012.00662.x)
- Fourie, J. (2014). The quantitative Cape: A review of the new historiography of the Dutch Cape Colony. *South African Historical Journal*, 66(1), 142–168. doi: [10.1080/02582473.2014.891646](https://doi.org/10.1080/02582473.2014.891646)
- Fourie, J., & Green, E. (2018). Building the Cape of Good Hope Panel. *History of the Family*, 23(3), 493–502. doi: [10.1080/1081602X.2018.1509367](https://doi.org/10.1080/1081602X.2018.1509367)
- Fourie, J., & Swanepoel, C. (2018). 'Impending ruin' or 'remarkable wealth'? The role of private credit markets in the 18th-century Cape Colony. *Journal of Southern African Studies*, 44(1), 7–25. doi: [10.1080/03057070.2018.1403218](https://doi.org/10.1080/03057070.2018.1403218)
- Fourie, J., & van Zanden, J. L. (2013). GDP in the Dutch Cape Colony: The national accounts of a slave-based society. *South African Journal of Economics*, 81(4), 467–490. doi: [10.1111/SAJE.12010](https://doi.org/10.1111/SAJE.12010)
- Genealogical Institute of South Africa. (2008). *South African Genealogies*. Stellenbosch: Genealogical Institute of South Africa.
- Genealogical Institute of South Africa. (2012). *South African Families*. Stellenbosch: Genealogical Institute of South Africa.
- Gouws, N. B. (1987). The demography of whites in South Africa prior to 1820. *Southern African Journal of Demography*, 1(1), 7–15. Retrieved from https://hdl.handle.net/10520/AJA16824482_8
- Guelke, L., & Shell, R. (1983). An early colonial landed gentry: Land and wealth in the Cape Colony, 1682–1731. *Journal of Historical Geography*, 9(3), 265–286. doi: [10.1016/0305-7488\(83\)90183-4](https://doi.org/10.1016/0305-7488(83)90183-4)
- Guelke, L. (1988). The anatomy of a colonial settler population: Cape Colony 1657–1750. *The International Journal of African Historical Studies*, 21(3), 453–473. doi: [10.2307/219451](https://doi.org/10.2307/219451)
- Heese, J. A. (1971). *Die herkoms van die Afrikaner, 1657–1867*. Kaapstad: A. A. Balkema.

- Hollingsworth, T. H. (1968). The importance of the quality of the data in historical demography. *Daedalus*, 97(2), 415–432. Retrieved from <http://www.jstor.org/stable/20023820>
- Klancher Merchant, E., & Alter, G. (2017). IDS Transposer: A users guide. *Historical Life Course Studies*, 4, 59–96. doi: [10.51964/hlcs9339](https://doi.org/10.51964/hlcs9339)
- Malherbe, D. F. du T. (1966). *Family register of the South African nation*. Stellenbosch: Tegniek.
- Martins, I., Cilliers, J., & Fourie, J. (2019). Legacies of loss: The intergenerational outcomes of slaveholder compensation in the British Cape Colony. *Lund Papers in Economic History. Development Economics* (No. 197). Lund: Lund University, Department of Economic History. Retrieved from https://portal.research.lu.se/portal/files/61357218/LUPEH_197.pdf
- Mitford-Barberton, I., & White, V. (1969). *Some frontier families: Biographical sketches of 100 Eastern Province families before 1840*. Cape Town: Human and Rousseau.
- Neumark, S. D. (1957). *Economic influences on the South African frontier, 1652–1836*. Stanford, CA: Stanford University Press.
- Penn, N. (2014). Casper, Crebis and the knegt: Rape, homicide and violence in the eighteenth-century rural Western Cape. *South African Historical Journal*, 66(4), 611–634. doi: [10.1080/02582473.2014.925961](https://doi.org/10.1080/02582473.2014.925961)
- Piraino, P., Muller, S., Cilliers, J., & Fourie, J. (2014). The transmission of longevity across generations: The case of the settler Cape Colony. *Research in Social Stratification and Mobility*, 35, 105–119. doi: [10.1016/j.rssm.2013.08.005](https://doi.org/10.1016/j.rssm.2013.08.005)
- Rijpma, A., Cilliers, J., & Fourie, J. (2020). Record linkage in the Cape of Good Hope Panel. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(2), 112–129. doi: [10.1080/01615440.2018.1517030](https://doi.org/10.1080/01615440.2018.1517030)
- Ross, R. (1975). The 'White' population of South Africa in the eighteenth century. *Population Studies*, 29(2), 217–230. doi: [10.1080/00324728.1975.10410200](https://doi.org/10.1080/00324728.1975.10410200)
- Sadie, J. (2000). *The economic demography of South Africa* (Doctoral dissertation). Stellenbosch: Stellenbosch University. Retrieved from <http://hdl.handle.net/10019.1/51963>
- Simkins, C., & van Heyningen, E. (1989). Fertility, mortality, and migration in the Cape Colony, 1891–1904. *The International Journal of African Historical Studies*, 22(1), 79–111. doi: [10.2307/219225](https://doi.org/10.2307/219225)
- Swanepoel, C. (2017). *The private credit market of the Cape Colony, 1673-1834: An investigation into the role of wealth, property rights, and social networks* (Doctoral dissertation). Stellenbosch: Stellenbosch University. Retrieved from <http://hdl.handle.net/10019.1/100828>
- TANAP. (2010). *Towards a New Age of Partnership*. www.tanap.net
- TEPC project. (2008). *Transcription of Estate Papers at the Cape of Good Hope Project*. /z-wcorg/.
- van Duin, P., & Ross, R. (1987). *The economy of the Cape Colony in the 18th century*. Leiden: Centre for the History of European Expansion.
- van Leeuwen, M. H. D., & Maas, I. (2011). *HISCLASS: A Historical International Social Class Scheme*. Leuven: Leuven University Press.
- van Leeuwen, M. H. D., Maas, I., & Miles, A. (2002). *HISCO: Historical International Standard Classification Of Occupations*. Leuven: Leuven University Press.
- Willigan, J. D., & Lynch, K. A. (1982). *Sources and methods of historical demography*. Cambridge, MA: Academic Press.
- Zhoa, Z. (2001). Chinese genealogies as a source for demographic research: A further assessment of their reliability and biases. *Population Studies*, 55(2), 181–193. doi: [10.1080/00324720127690](https://doi.org/10.1080/00324720127690)

IV

Specific cohorts



HISTORICAL LIFE COURSE STUDIES
VOLUME 12 (2022), published 08-09-2022

Nominative Linkage of Records of Officials in the China Government Employee Dataset-Qing (CGED-Q)

Cameron Campbell

The Hong Kong University of Science and Technology & Central China Normal University & 2022–23 Fellow, Center for Advanced Study in the Behavioral Sciences, Stanford University

Bijia Chen

Renmin University

ABSTRACT

We introduce our approach to the nominative linkage of records of Qing officials who were included in the China Government Employee Datasets-Qing (CGED-Q) Jinshenlu (JSL) and Examination Records (ER). We constructed these datasets by transcription of quarterly rosters of civil and military officials produced by the government and by commercial presses, and records of examination degree holders. We assess each of the primary attributes available in the original sources in terms of their usefulness for disambiguation, focusing on their diversity and potential for inconsistent recording. For officials who were not affiliated with the Eight Banners, these primary attributes include surname, given name, and province and county of origin. For the small subset of officials who were affiliated with the Bannermen, we assess the available data separately. We also assess secondary attributes available in the data that may be useful for adjudicating candidate matches. We then describe the approach that we developed that addresses the issues we identified with the primary and secondary attributes. The issues we have identified and the approach that we have developed will be of interest to researchers engaged in similar efforts to construct and link datasets based on elite males in historical China.

Keywords: China, Nominative linkage, Elites, Careers

DOI article: <https://doi.org/10.51964/hlcs11902>

© 2022, Campbell, Chen

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

We describe our approach to the large-scale nominative linkage of records of elite males in two Qing dynasty (1644–1911) historical datasets that we have constructed: the China Government Employee Dataset-Qing Jinshenlu (CGED-Q JSL) and Examination Records (CGED-Q ER). By transcribing records of Qing civil and military officials in quarterly personnel rosters from the period between 1762 and 1911 to produce the CGED-Q JSL and then linking those records over time, we have reconstructed the career histories of officials. By linking officials in the CGED-Q JSL to their records in the CGED-Q ER, we have also attached information about their year of birth, exam performance, ancestry, and other attributes to their career records. This allows us to examine a major topic in the sociological study of stratification: the roles of family background and 'ability' (as measured by exam performance) in the appointment, promotion, and exit from work of officials. Attaching information on year of birth to career histories allows for the study of the age structure of officialdom and the age dynamics of appointment, promotion, and exit from work.

We arrived at the approach we describe here iteratively, building on experience analyzing career histories in the CGED-Q JSL in a series of publications on appointment, promotion, and exit of Qing officials (Campbell, 2020; Chen, Campbell, & Lee, 2018; Hu, Chen, & Campbell, 2020; Hu, Hu, Chen, & Campbell, 2021; Xue & Campbell, 2022), a visualization platform (Wang et al., 2021), an introduction to the CGED-Q JSL (Chen, Campbell, Ren, & Lee, 2020) and a dissertation (Chen, 2019). Each analysis brought to light issues with the sources, the transcription process, and linkage procedures that had not arisen previously and required adjustments. As our dataset expanded, meanwhile, we adjusted our code and obtained substantial improvements in speed. In the end, as described below, we used probabilistic linkage as implemented in the STATA package *dtalink* (Kranker, 2018).

The most important contribution of the paper is the thorough documentation of the many problems that arise in the recording of names, place of origin, and other attributes in Qing administrative sources, the implications of these problems for nominative linkage, and our solutions to them. We hope that our experience will be useful to researchers carrying out large-scale nominative linkage in other Chinese sources and to users of the CGED-Q JSL public releases that we have made available for download (Campbell, Chen, Ren, & Lee, 2019).¹ The problems that we identify and our solutions to them should be general to historical Chinese sources. Common problems include the replacement of characters in surnames and given names with variant forms, homonyms, and similar-looking characters, and the inconsistency in the recording of locations because of changes in administrative boundaries. To facilitate work by others who are carrying out nominative linkage with historical Chinese sources, we have also made the complete tabulations that are the basis of most of our tables available for download.²

We organize our paper as follows. Section 2 presents a brief review of the literature on nominative linkage in historical and Chinese language sources. Sections 3 and 4 introduce the two datasets that we link: the China Government Employee Dataset-Qing Jinshenlu (CGED-Q JSL) and the Examination Records (CGED-Q ER). We describe the attributes of officials recorded in the data that may be used for linkage, distinguishing between primary attributes available for all officials in both sources, and secondary attributes only available in the CGED-Q JSL. We identify issues that arise in the primary attributes that need to be addressed when carrying out linkage. Section 5 describes our current approaches to linkage in the CGED-Q JSL and CGED-Q ER. Section 6 concludes with discussion of implications of these results and prospects for the future.

1 We have released CGED-Q JSL data for the years 1850–1864 and 1900–1912. The data and documentation may be downloaded at the Lee-Campbell Group page at the HKUST Dataspace (<https://doi.org/10.14711/dataset/E9GKRS>) and the Lee-Campbell Group page at the Harvard Dataverse (<https://doi.org/10.7910/DVN/GMQWVZ>). We will release additional tranches of CGED-Q JSL data every two years until we have made it all available, and then release the CGED-Q ER.

2 As a resource for other researchers carrying out nominative linkage with historical Chinese sources, we have made the complete tabulations that were the basis of tables 2 through 8 available for download at the Lee-Campbell Group dataverses at the HKUST Dataspace (<https://doi.org/10.14711/dataset/M8HQEA>) and at the Harvard Dataverse (<https://doi.org/10.7910/DVN/4OSP8V>). These should help researchers who need to address issues related to inconsistency in the recording of names or places of origin develop approaches to handle such problems.

2 BACKGROUND

Large-scale automated nominative linkage of records of individuals from archival sources is a key tool for production of longitudinal 'big data' for ongoing studies of European and North American population, social and economic history. Common applications are the linkage of census records for the same individual at different points in time, and linkage of individuals across birth, death, and marriage records. Linkage may also include other, more specialized sources including tax records, health records, and retirement and pension records that supplement information routinely available in census records and vital registration. The resulting linked data not only provide life histories of individuals, but in some cases, histories of families across multiple generations. As a result of this activity, methods for large-scale nominative linkage of individuals in sources written in English and other languages that use phonetic scripts are relatively mature. A large literature discusses challenges associated with nominative linkage and offers various solutions, and relevant software packages are readily available.³

The literature on large-scale nominative linkage of records of individuals with names written in English and other phonetic scripts in historical sources is already large because efforts to construct massive longitudinal databases for social and economic history by linkage of censuses, vital records and other administrative data have been underway in the United States, Canada, and a variety of European countries for at least two decades.⁴ An early example was the initiative by the Minneapolis Population Centre to create a statistically representative sample of records in the 1860, 1870, 1900 and 1910 censuses linked to the complete-count 1880 census described in Ruggles (2002). Since then, methodology has advanced substantially, with explorations of machine learning to fully automate record linkage (Abramitzky, Mill, & Pérez, 2020) and the leveraging of information on residence and relationships to increase linkage rates (Akgün et al., 2020; Helgertz et al., 2020).

Key issues that arise in the linkage of names written in English and other languages with phonetic alphabets include misspellings, name changes, the use of variant spellings, and inconsistencies in the recording of other attributes like age or date of birth, all of which could create false negatives, and overall low diversity of surnames and given names, which could lead to false positives. By 'false negatives', we refer to situations where two records that should have been linked together were not. By 'false positives', we refer to situations where records that should not have been linked together, were. Misspellings occurred because people were inconsistent in the way they wrote their own name, or the way census takers or other officials wrote their name in official records. International migrants might have new names assigned to them by immigration officers who transliterated their original names in ad hoc fashion or might adapt new names on their own. Women typically adopted their husband's surname on marriage. People might use contractions of their name or nicknames in some situations but not in others, for example, writing Bill in some situations and William in others. In many communities in Europe, diversity of surnames and given names was low, making it difficult to distinguish whether records of the same name referred to the same or different people.

The issues that arise with names written in Chinese are very different. Surnames are not diverse. In 2020, the top 5 surnames in China accounted for 30.8% of the population, and the top 100 surnames accounted for 85.8% of the population.⁵ Given names are potentially more diverse since they are typically two characters, and for each of those two characters there are thousands to choose from. The actual diversity of given names depended on naming practices in different periods and social classes. While names of elite males during the Qing and the first half of the 20th century should have been very diverse because well-off families could showcase their erudition by including rare characters with literary, historical or philosophical connotations in the names of their sons, names for people born between the 1960s and 1980s were much less diverse than for those born before or after because

3 See, for example, the Linkage Library at <https://www.icpsr.umich.edu/web/pages/about/linkage-library.html>.

4 See *Historical Methods* Special Issues 51(2) and 53(4) on historical record linkage for introductions to relevant projects (Sylvester & Hacker, 2020).

5 See the 2019 and 2020 *Nian Quanguo Xingming Baogao* [National Surname and Given Name Report] published by the Public Security Bureau of the People's Republic of China, retrieved from http://www.gov.cn/xinwen/2021-02/08/content_5585906.htm.

single character names with political or patriotic implications became more popular (Cai, Xi, Yi, Liu, & Jing, 2018; Bao, Cai, Jing, & Wang, 2021).⁶

Developing procedures for record linkage is important because there are numerous efforts ongoing to create biographical databases of historical Chinese individuals. Prominent examples include the China Biographical Database (Chen & Wang, 2022; Fuller, 2021; Tsui & Wang, 2020), the Modern China Historical Database (Armand, Guo, Henriot, Hu, & Van den Bosch, 2022), and the various projects of the Lee-Campbell Group (Campbell & Lee, 2020). Such databases are the basis of prosopographical studies of social groups (Stone, 1971), especially elites, in historical China. The creators of these databases carry out what they refer to as 'disambiguation' to assess whether the same name and other attributes appearing in two or more sources refers to the same person or different people, and then attach unique identifiers to each appearance of a person in the dataset. The underlying task is similar to the record linkage that we carry out in the CGED-Q, but somewhat broader in that it may also involve individuals named in unstructured texts like newspaper articles, dynastic histories, or gazetteers.⁷ Pronunciation-based approaches developed for linkage of individuals with names written in phonetic scripts are not immediately useful for these Chinese language sources because the prevalence of homonyms in Chinese means that names with identical pronunciations can be completely different. Meanwhile, characters that look similar and may be mistakenly replaced with each other during the production process can be pronounced differently and have different meanings.

The studies of Chinese language nominative linkage and disambiguation that we have located focus on names in contemporary unstructured Chinese language texts (for example, web pages) not on structured records like in the CGED-Q. We mention them here because they could eventually help with the linkage of the officials in the CGED-Q to mentions of them in unstructured texts. Chen and Huang (2010) assessed issues that arise in the disambiguation of the names of individuals in Chinese language texts. They report that single character given names are more challenging than two character given names. Combinations of surname and single-character name that are also commonly used words are especially difficult to disambiguate. For example, the combination *Gaofeng* (高峰) could be the surname Gao followed by the given name Feng but could also be the word for 'peak'.⁸ Han, Zu and Zhao (2011) and Fan and Li (2021) describe approaches based on clustering in which the same names appearing in different documents are disambiguated by reference to other words appearing with them in the text. The problems these papers address is different to the one we face in our own linkage of names in tabular datasets where the surname and given name are clearly specified in fields of their own, but relevant for efforts by others to extract and disambiguate names in unstructured historical texts like newspaper articles, books, and essays.

Several studies discuss the disambiguation of Chinese names of authors of texts. Han et al. (2017) focus on the specific case of disambiguating the names of authors of Chinese language publications, and introduce a method based on the names of the co-authors, the author's institution, and 'semantic fingerprints'. Kim, Kim and Kim (2021) shows that disambiguation of the names of Chinese authors of English language publications is easier if their name in Chinese characters is available alongside their phoneticized names. Yin, Motohashi and Dang (2020) presents the results of an effort to disambiguate the names of inventors listed on Chinese patents between 1985 and 2016. They use supervised learning approach that begins with hand-labelled data for training.

Another line of studies offers potentially useful approaches for measuring similarity in the sound and the appearance of Chinese characters and then using this to assess the similarity of strings of Chinese characters. Liu, Rus, Liao and Liu (2017) offer a method for encoding Chinese characters in terms of their sound,

6 See Chua (2021) for an overview of contemporary naming practices in China and descriptive results on the popularity of different kinds of names during the 20th century. The analysis was based on the Chinese Name Database (1930–2008) created by Han-Wu-Shang (Bruce) Bao and shared at <https://github.com/psychbruce/ChineseNames>.

7 Campbell and Lee (2020) and Chen and Campbell (2023) include brief, non-technical overviews of linkage in the CGED-Q as part of their overviews of the methods used in the project. We describe linkage procedures for two of our other publicly released datasets, the China Multigenerational Panel Datasets (CMGPD) Liaoning (LN) and Shuangcheng (SC), in Appendix A of Lee and Campbell (1997), Lee, Campbell and Chen (2010) and Wang et al. (2013). According to personal communication with the leaders of the China Biographical Database and Modern China Historical Database projects, they do not yet have any publications describing their procedures for linkage and disambiguation.

8 Segmentation of text is also important because in the absence of spaces between words, there are instances where the last character of one word and the first character of the word that immediately follows might be mistaken for a name.

appearance, and meaning, and then ranking pairs of characters according to their similarity. Chen et al. (2018) proposes a "SoundShape Code" for Chinese characters that reflects their pronunciation and appearance, and which may be used as a basis of measuring similarity between two characters in a pair. Xu, Zheng and Li (2020) combine the SoundShape Code for individual Chinese characters with the Dice similarity measure for strings of potentially different lengths. Such methods address a challenge that we describe below: because of errors in the original source or errors during our transcription, the names of the same individual may appear with slightly different characters in different records in our dataset. Characters may be replaced by a homonym that looks different, or with a visually similar character that is pronounced very differently.

3 CHINA GOVERNMENT EMPLOYEE DATASET-QING JINSHENLU (CGED-Q JSL)

We constructed the China Government Employee Dataset-Qing Jinshenlu (CGED-Q JSL) from *Jinshenlu* (縉紳錄) and *Zhongshubeilan* (中樞備覽) rosters of Qing civil and military officials respectively that were produced every three months. We have described the CGED-Q JSL and the sources from which it was constructed in detail elsewhere (Chen et al., 2020; Ren, Chen, Hao, Campbell, & Lee, 2016, 2019) and only provide key details here. Official editions of the *Jinshenlu* and *Zhongshubeilan* were produced by the Qing Ministries of Personnel and War, respectively.⁹ The government used the official editions to keep track of posts and the officials who held them. In the 19th century, commercial publishers produced and sold editions that supplemented information on officials from the official editions with additional information collected by the publishers.¹⁰ Purchasers of commercial editions used them for a variety of purposes, including searching for vacant positions and locating kin, classmates, or other connections who they knew were officials.

At the time of writing, the CGED-Q JSL contains 4,433,600 records from 275 *Jinshenlu* editions and 75 *Zhongshubeilan* editions. Each *Jinshenlu* roster lists 13,000 to 15,000 posts in the civil service and identifies the officials who held them. *Zhongshubeilan* rosters each list approximately 8,000 military posts and the officers who held them. The editions in the CGED-Q JSL are from the period 1762 to 1912. Coverage is sparse before 1830, but very complete after that year. From 1830 to 1911, the CGED-Q JSL includes at least one *Jinshenlu* edition from nearly every year. In many years, it includes all four quarterly editions. *Zhongshubeilan* are sparser and the gaps between them are longer.

78.9% of officials were ordinary citizens (*minren*, 民人) and almost all the remainder were Bannermen (*qiren*, 旗人). The vast majority of *minren* were what we would now refer to as Han Chinese.¹¹ Bannermen were hereditary affiliates of the Eight Banners, originally the army used to conquer China and establish the Qing in 1644, and in the 18th and 19th centuries, an organization used by the Qing state to maintain political and military control. Most officials who were Bannermen were Manchu or Mongol, but 16.4% were Han Chinese. The latter were referred to as Han Martial Bannermen (*hanjun qiren*, 漢軍旗人). They were the descendants of Han Chinese who had been incorporated into the Eight Banners. Bannermen had a privileged position in the Qing government, with their own pathways to appointment and promotion, and quotas for certain positions. Thus, even though Bannermen accounted for only 2%–4% of the population of the Qing (Elliott, Campbell, & Lee, 2016), they accounted for one-fifth of civil officials overall, two-thirds of civil officials serving in the capital Jingshi (now Beijing) and 90% of officials in the secondary capital Shengjing (now Shenyang) (Chen et al., 2020, p. 454).

To produce career histories by longitudinal linkage of CGED-Q JSL records of officials, we distinguish between what we refer to as the primary and secondary attributes recorded for officials. We define primary

9 We refer to the dataset constructed from *Jinshenlu* and *Zhongshubeilan* rosters as CGED-Q Jinshenlu (JSL) because when *Zhongshubeilan* are available, it is usually as part of a set with a *Jinshenlu* edition for the same season. The resulting sets are typically catalogued by libraries and archives as a *Jinshenlu* edition. We only have a small number of freestanding *Zhongshubeilan* editions that are not part of a set.

10 See Chen et al. (2020) for a detailed discussion of the differences in the contents of the official and commercial editions.

11 What we refer to as *minren* likely also included members of what since the 1950s have been officially designated as minority ethnic groups, but the *Jinshenlu* does not record any information that would allow us to distinguish them.

attributes as basic and stable information about an official that are available in all or nearly all records and should be available in almost any other source that we might wish to link to. The most important of these are the names. We define secondary attributes as characteristics that are specific to the CGED-Q JSL and may not be available in other sources or recorded in every edition of a *Jinshenlu* or *Zhongshubeilan*. They may also be attributes that vary over time, for example, the official's current position. These may be used to adjudicating candidate links made based on the primary attributes, but on their own are not sufficient for linkage within the CGED-Q JSL or between the CGED-Q JSL and CGED-Q ER.

For linkage, we separate officials according to whether they had a surname recorded because the primary attributes available for officials with surnames differed from those available for those without surnames. Officials with surnames accounted for 80.2% of records. These included all the *minren* and one-third of the Han Martial Bannermen.¹² Basic information recorded for them included not only their surname (*xing*, 姓) and given name (*ming*, 名) but also their place of origin. The latter was usually the province and prefecture or county of origin, though there are complications that we discuss below. Officials without surnames included all Manchu (*Manzhou*, 滿洲) and Mongol (*Menggu*, 蒙古) Bannermen and two-thirds of Han Martial Bannermen.¹³ The only attributes recorded for officials without surnames that were in principle stable were given name and Banner affiliation (*qifen*, 旗分). We use these as the primary attributes for Bannermen.

The primary attributes for officials with and without surnames differ in terms of their ability to uniquely identify officials within an edition. For officials with a surname, the combination of surname, given name, and province and county of origin was usually unique within an edition. If these were all recorded reliably and consistently across every edition, they would in principle be sufficient for linkage. Table 1 summarizes the number of repetitions of combinations of primary attributes within each quarterly *Jinshenlu* edition. For officials with a surname, 95.0% of the combinations of surname and given name were unique within their edition. In other words, for 95.0% of records, there was no other record in the same edition with the same surname and given name. For 4.4% of records, there was only one other record in the same edition with the same surname and given name. 98.1% of records of officials with a surname were unique within their edition in terms of the combination of surname, given name, and place of origin. Our investigations have revealed that for these officials, most repetitions within the same edition all refer to the same official. If an official held more than one post, there was a separate record for each of them.

Table 1 *Uniqueness of primary attributes of officials within each quarterly edition of the Jinshenlu, 1760–1912*

	Officials with a surname		Officials without a surname	
	Surname and Given name	+ Place of origin	Given name	+ Banner
Repetitions within an edition ^a	%	%	%	%
1	95.0	98.1	64.0	88.0
2	4.4	1.7	19.9	9.9
3	0.5	0.2	8.2	1.4
4	0.1	0.0	4.0	0.4
5 or more	0.01	0.0	3.9	0.4
Total	100	100	100	100
Records	2,817,156	2,817,156	784,502	784,502

^a *Repetitions* refers to the total number of records in the same quarterly edition with the specified combination primary attributes.

For officials without surnames, given name by itself is not sufficient for linkage. Only two-third of records recorded a given name that was unique within the quarterly edition. One-third of records had a name that appeared in one or more other records. When Banner affiliation was added, 88% of records became unique within their quarterly edition in terms of the primary attributes. 12% of records had a given name and Banner affiliation that appeared in at least one other record in the same

12 Exactly one-third of the officials recorded as Han Marital Bannermen had a surname recorded. The remainder did not, presumably because they had taken Manchu names. See Campbell, Lee and Elliott (2002) for a discussion of the adaptation of Manchu names by Han Chinese in northeast China.

13 Manchu and Mongol Bannermen accounted 71.4% and 12.2% of Bannermen, respectively.

quarterly education. Based on our investigations, these reflect some cases where the same official held more than one office, as well as cases where two different officials had the same name.

These results highlight that the approaches to linkage must differ according to whether a surname was available. For officials with a surname, as discussed above, the combination of surname, given name, and province and county of origin all written in Chinese characters is likely to be unique, and false positives in which records of different officials are mistakenly linked together should be rare. The main task for linkage of officials with a surname is avoiding false negatives in which an inconsistency in the recording of the name or some other attribute prevents a link from being made. For officials without a surname, the risk of false positives is high because surnames and place of origin are not available, and there are enough officials who share the same combination of given name and Banner affiliation to raise concerns that two records with identical name and Banner affiliation may refer to different officials.

Below we introduce the primary and secondary attributes in detail and assess their usefulness for linkage, with a focus on their homogeneity or heterogeneity. We divide our discussion of attributes between those available in records of officials with surnames and those available in records of officials without surnames.

3.1 ATTRIBUTES AVAILABLE FOR OFFICIALS WITH SURNAMES

3.1.1 SURNAMES

Because a small number of surnames accounted for a large share of the records of officials, surnames are of limited utility as a primary attribute for linkage. According to Table 2, which presents the cumulative percentages of records accounted for by the 100 most common surnames in the CGED-Q JSL, the five most common surnames appeared in one-quarter of the records. These were Wang (王), Zhang (張), Li (李), Chen (陳) and Liu (劉). The top 10 surnames accounted for 38.3% of the records of officials with surnames. The top 20 surnames accounted for approximately one-half of the records, and the top 200 accounted for 95.1%. There were a total of 1626 distinct surnames recorded, though the actual number was lower because in this tabulation a surname may have more than one entry if the character appears in more than one form.

Table 2 *Cumulative percentages of the top 100 most common surnames in the CGED-Q JSL, 1760–1912*

	1–20		21–40		41–60		61–80		81–100	
	Surname	%	Surname	%	Surname	%	Surname	%	Surname	%
1	王	6.6	林	51.6	蔡	65.4	魏	75.0	薛	81.5
2	張	12.7	謝	52.4	韓	65.9	戴	75.4	廖	81.8
3	李	18.7	郭	53.3	唐	66.5	盧	75.7	白	82.0
4	陳	23.5	高	54.1	鄧	67.1	田	76.1	嚴	82.3
5	劉	27.7	許	54.9	蔣	67.6	崔	76.5	萬	82.6
6	楊	30.6	馮	55.6	方	68.2	夏	76.8	施	82.8
7	周	32.8	吳	56.4	孔	68.7	熊	77.2	賈	83.1
8	吳	34.7	羅	57.1	蕭	69.3	陶	77.5	洪	83.3
9	徐	36.5	梁	57.8	袁	69.8	秦	77.8	雷	83.6
10	趙	38.3	姚	58.5	曾	70.3	俞	78.2	邱	83.8
11	朱	40.0	葉	59.2	董	70.8	江	78.5	姜	84.1
12	孫	41.5	程	59.9	章	71.3	譚	78.8	孟	84.3
13	胡	42.9	余	60.5	傅	71.7	鄒	79.2	賀	84.5
14	馬	44.2	宋	61.1	錢	72.2	史	79.5	毛	84.8
15	沈	45.4	潘	61.7	顧	72.6	于	79.8	侯	85.0
16	黃	46.6	丁	62.4	范	73.0	鍾	80.1	尹	85.2
17	何	47.8	彭	63.0	杜	73.4	龔	80.4	武	85.4
18	鄭	48.8	陸	63.6	蘇	73.8	邵	80.7	郝	85.6
19	黃	49.7	曹	64.2	任	74.2	石	80.9	葛	85.8
20	汪	50.7	金	64.8	呂	74.6	湯	81.2	倪	85.9

Note: Based on authors' calculations on 3,244,484 CGED-Q JSL records with a legible surname.

One issue that arises with linkage based on surnames is that a character may be replaced with one that looks similar in an adjacent edition. Of the 1,559,380 pairs of records in editions which were no more than one year apart and almost certainly referred to the same official because they recorded the same two-character given name, province and county of origin, and position and broad category of degree qualification, 20,055 pairs (1.3%) differed on the character written for the surname.¹⁴ Table 3 presents the cumulative frequencies of discordant pairs of surnames. The most common discordant pair (黃 黃) accounted for 22.4% of discordant pairs overall, the top 20 accounted for nearly two-thirds (63.1%) and the top 100 accounted for 79.2%.

Inspection of the results in Table 3 reveals two common issues that may generate false negatives, in which records of the same official are not properly linked. The first issue is that some pairs are the same character written in variant forms (*Yitizi*, 異體字). The four most common pairs in Table 3 are examples: 黃 and 黃, 吳 and 吳, 高 and 高, and 呂 and 呂 are different ways of writing the surnames Huang, Wu, Gao, and Lu respectively. In the Unicode standard these are recognized as different representations of the same character, and as we describe below, this is straightforward to address. The second and more challenging issue is that sometimes between editions a character for a surname is replaced by one that looks similar but is a completely different character. Examples in Table 3 include the fifth entry (段, Xia and 段, Duan), the seventh entry (宋, Song and 朱, Zhu), the 10th entry (汪, Wāng and 王, Wáng), and the 15th entry (馬, Ma and 馮, Feng). These issues reflect either inconsistencies in the production process across different editions or transcription errors by coders. There are also examples of discordant pairs in Table 3 that consist of characters that are clearly different, for example the 24th entry 張 章 (Zhang and Zhang) and the 28th entry 程 陳 (Cheng and Chen). In most of these cases, one or both characters are relatively common surnames. While there is some possibility that these could be from records of different people, they may also be transcription errors that occurred during data entry.

Table 3 Cumulative percentages of the top 100 most common discordant pairs of surnames in adjacent editions in the CGED-Q JSL

	1–20		21–40		41–60		61–80		81–100	
	Pair	%	Pair	%	Pair	%	Pair	%	Pair	%
1	黃黃	22.4	衛衛	63.7	鄧鄭	70.7	盧虞	74.6	蔣薛	77.2
2	吳吳	35.7	關關	64.2	曹曾	71.0	丁于	74.7	翰韓	77.3
3	高高	41.4	孫馮	64.7	章童	71.2	閔關	74.9	向尚	77.4
4	呂呂	44.8	張章	65.1	杜林	71.5	馮馮	75.0	俞喻	77.5
5	段段	47.2	柳柳	65.6	余徐	71.7	葉蔡	75.1	褚諸	77.6
6	錢錢	49.3	劉陳	66.0	徐涂	71.9	曾魯	75.3	徐許	77.7
7	宋朱	51.3	甯甯	66.4	全金	72.1	張陳	75.4	束束	77.8
8	閆閆	52.8	程陳	66.8	鄔鄔	72.4	董黃	75.5	寇寇	78.0
9	汪王	54.1	楊陽	67.1	董黃	72.6	刑邢	75.7	樂樂	78.1
10	凌凌	55.3	余金	67.5	員貧	72.8	宋宗	75.8	張楊	78.2
11	賴賴	56.5	楊湯	67.8	于王	72.9	萬黃	75.9	苑范	78.3
12	余俞	57.5	毛王	68.1	李陳	73.1	強強	76.1	郭鄧	78.4
13	龐龐	58.4	余余	68.4	吳呂	73.3	王黃	76.2	婁婁	78.5
14	溫溫	59.2	寶寶	68.8	晉晉	73.5	曹曹	76.3	柏栢	78.6
15	馬馮	59.9	季李	69.1	曹賈	73.6	潘王	76.4	丁李	78.7
16	涂涂	60.6	嵇稽	69.4	童董	73.8	湛湛	76.6	褚褚	78.8
17	顏顏	61.2	龍龔	69.6	劉鄧	74.0	杜樊	76.7	範範	78.9
18	閔關	61.9	侯候	69.9	邊邊	74.1	唐康	76.8	廉廉	79.0
19	江汪	62.5	朱李	70.2	瞿翟	74.3	寇寇	76.9	呂吳	79.1
20	鍾鐘	63.1	陳陸	70.5	宮宮	74.4	荆荆	77.0	孫張	79.2

Note: Of the 1,559,380 pairs of records in adjacent editions no more than one year apart that were identical on given name, province and county of origin, broad category of degree qualification, and position, 20,055 (1.3%) were discordant.

14 Degree qualification refers to the examination or purchased degree that qualified a *minren* official for appointment to office. This is a secondary attribute that we discuss below.

3.1.2 GIVEN NAMES

Given names (Table 4) were the most diverse of the primary attributes available for officials with surnames, and therefore the most useful for record linkage. We distinguish between records of officials with two- and one-character names. The former accounted for 85% of the records and the latter accounted for the remainder. A total of 102,648 distinct given names appeared in our data, 98,745 of which were two-character names, with the remaining 3,903 being one-character names. According to Table 4, two-character names were very diverse. The top 100 accounted for only 5.7% of records, the top 200 accounted for 9% of records, the top 1,000 accounted for 23% of records, and the top 10,000 accounted for only 61% of records. The diversity of two-character names reflects the large number of characters available to choose from: we found that at least 5,764 different characters made at least one appearance in a two-character given name in the CGED-Q JSL.¹⁵

Table 4 Cumulative percentages of the top 100 most common two-character given names of officials with surnames in the CGED-Q JSL

	1–20		21–40		41–60		61–80		81–100	
	Given name	%	Given name	%	Given name	%	Given name	%	Given name	%
1	汝霖	0.1	樹棠	1.7	瑞麟	2.9	祖培	3.9	錫麟	4.9
2	文炳	0.2	炳文	1.8	桂芳	3.0	繼昌	4.0	登雲	4.9
3	得勝	0.3	雲龍	1.8	殿元	3.0	沛霖	4.0	文彬	4.9
4	占魁	0.4	桂林	1.9	玉麟	3.1	祖蔭	4.1	安邦	5.0
5	兆麟	0.5	占鰲	2.0	國泰	3.1	鴻鈞	4.1	錫疇	5.0
6	作霖	0.6	逢春	2.0	維藩	3.2	其昌	4.2	建勳	5.1
7	廷棟	0.7	廷桂	2.1	恩培	3.2	鵬飛	4.2	鴻恩	5.1
8	秉鈞	0.8	鳳翔	2.2	紹曾	3.3	炳章	4.3	毓麟	5.2
9	承恩	0.9	步雲	2.2	文蔚	3.3	炳南	4.3	玉堂	5.2
10	慶雲	1.0	國楨	2.3	殿魁	3.4	國祥	4.4	樹森	5.2
11	世昌	1.0	煥章	2.3	桂森	3.4	長庚	4.4	念祖	5.3
12	步瀛	1.1	文藻	2.4	國華	3.5	定邦	4.4	桂芬	5.3
13	兆熊	1.2	長春	2.5	光祖	3.5	振邦	4.5	學海	5.4
14	培元	1.2	登瀛	2.5	國瑞	3.6	萬春	4.5	連陞	5.4
15	文光	1.3	慶元	2.6	廷珍	3.6	慶恩	4.6	家駒	5.4
16	維翰	1.4	維城	2.6	世榮	3.7	永清	4.6	錫祺	5.5
17	樹勳	1.4	恩榮	2.7	恩溥	3.7	永清	4.7	文治	5.5
18	文煥	1.5	錫齡	2.7	維新	3.8	廷杰	4.7	濟川	5.6
19	錫恩	1.6	國棟	2.8	春華	3.8	榮光	4.8	占春	5.6
20	振聲	1.6	壽昌	2.8	遇春	3.9	廷楨	4.8	鶴年	5.7

Note: Based on authors' calculations on 2,718,433 CGED-Q JSL records with a legible surname and a legible two-character given name.

Like surnames, characters in given names may also be inconsistent across different quarterly editions. If not addressed, this may also lead to false negatives. Table 5 repeats the exercise for surnames carried out in Table 3 for the characters in two-character given names.¹⁶ It presents the cumulative percentages of discordant pairs, defined as characters in given names that differ between records in editions that are no more than one year apart, and where the surname, one of the two characters in the given name, place of origin, position, and degree qualification are all identical. Out of 1,539,198 such pairs of records, 4.34% (66,994) differed on one character in the given name. Discordant pairs of characters in given names were much more diverse than was the case for surnames. The most common

15 We have included the complete tabulation of characters making at least one appearance in a two-character given name as one of the files available for download at the Harvard and HKUST Dataverses for this paper.

16 Restricting to a two-character name and then including the requirement that at least character in the name matches substantially increases the likelihood that two records that match on everything else refer to the same person.

discordant pair (清 and 淸) accounted for only 3.7% of discordant pairs. The top 20 accounted for one-fifth (20.3%) of discordant pairs, and the top 100 accounted for 39.2%.

Once again, the most common issue is that between one edition and the next, a character was replaced with a variant, of which the seven most frequent pairs are all examples. 清 and 淸, for example, are both ways of writing the same character (Qing). However, there are also cases where a character is replaced by one that is different but looks similar. The 12th, 14th, 22nd and 39th entries are examples: 傳 (Fu) and 傳 (Chuan), 思 (Si) and 恩 (En), 增 (Zeng) and 曾 (Ceng), and 先 (Xian) and 光 (Guang), respectively. Again, this likely reflects a problem during the production of the source, or during the transcription.

Single-character names were less diverse. According to Table 6, the top 10 most common single-character names accounted for 6.6% of records with single-character names and the top 100 accounted for 37%. According to separate tabulations, the top 200 accounted for 54% and the top 500 accounted for 78%. According to a separate tabulation like the ones in Tables 3 and 5 but not shown here, the patterns in discordant pairs are like those in Table 5. Most discordant pairs consisted of the same characters written differently or similar looking characters that could be mistaken for each other. There were examples, however, of characters that were clearly different, at least raising the possibility that they were men from the same county with the same surname and post who should not be linked. Accordingly, we link records of officials with single-character names separately, with more stringent criteria for match on other attributes when assessing candidate links.

Table 5 Cumulative frequencies of the 100 most common discordant pairs of characters in two-character given names in records of officials with surnames in adjacent editions in the CGED-Q JSL

	1-20		21-40		41-60		61-80		81-100	
	Pair	%	Pair	%	Pair	%	Pair	%	Pair	%
1	清清	3.7	鳴鳴	20.8	覲覲	27.8	元光	32.7	得德	36.4
2	勳勳	5.7	增曾	21.2	之芝	28.1	堯堯	32.9	台臺	36.6
3	齡齡	7.0	曾會	21.6	穀穀	28.4	峯峰	33.1	宜宣	36.7
4	鳳鳳	8.3	遠遠	22.0	春椿	28.6	榮榮	33.3	城成	36.9
5	壽壽	9.5	延延	22.4	壁壁	28.9	世士	33.5	捷捷	37.0
6	寶寶	10.7	廉廉	22.8	緒緒	29.2	寬寬	33.7	嘉家	37.2
7	晉晉	11.7	懷懷	23.2	變變	29.4	顯顯	33.9	彝彝	37.3
8	煥煥	12.7	耀耀	23.6	凌凌	29.7	為為	34.1	日日	37.5
9	賓賓	13.6	慎慎	24.0	瀚翰	29.9	惟維	34.3	如汝	37.6
10	彥彥	14.4	濂濂	24.3	繩繩	30.1	甲申	34.5	連運	37.8
11	恆恆	15.2	熙熙	24.7	葆葆	30.4	輝輝	34.7	宗崇	37.9
12	傳傳	15.9	猷猷	25.0	崧松	30.6	昭照	34.8	誠誠	38.1
13	青青	16.7	瀾瀾	25.4	均鈞	30.9	恩榮	35.0	彌彌	38.2
14	思恩	17.3	蔡芬	25.7	柱桂	31.1	瑞端	35.2	燿燿	38.4
15	鍾鐘	17.9	蕃藩	26.0	聯聯	31.3	祿祿	35.4	柏栢	38.5
16	鎮鎮	18.5	高高	26.3	豐豐	31.6	繩繩	35.6	彝彝	38.6
17	庭廷	19.0	啓啟	26.7	方芳	31.8	丙炳	35.7	讓讓	38.8
18	熙熙	19.4	樹澍	27.0	迪迪	32.0	璋章	35.9	鰲鰲	38.9
19	達達	19.9	先光	27.3	祐祐	32.3	堂堂	36.1	萼萼	39.1
20	聯聯	20.3	聯聯	27.6	舉舉	32.5	清青	36.2	達達	39.2

Note: In the 1,539,198 pairs of records with legible surname and two-character given names in adjacent editions no more than one year apart that were identical on surname, one character of the given name, province and county of origin, broad category of degree qualification, and position, there were 66,994 discordant pairs.

Table 6 Cumulative percentages of the top 100 most common one-character given names of officials with surnames in the CGED-Q JSL

	1–20		21–40		41–60		61–80		81–100	
	Character	%	Character	%	Character	%	Character	%	Character	%
1	鈞	0.9	芳	12.0	煜	20.2	璋	26.9	溶	32.5
2	榮	1.7	煦	12.4	勳	20.6	煥	27.2	琦	32.7
3	鑑	2.4	淦	12.9	潤	21.0	桐	27.5	瑛	33.0
4	炳	3.0	源	13.4	濤	21.3	鎔	27.8	坦	33.2
5	鏞	3.7	灃	13.8	鵬	21.7	玉	28.1	超	33.5
6	鈺	4.3	浩	14.3	鴻	22.0	筠	28.4	鎬	33.7
7	瀛	4.9	培	14.7	沅	22.4	治	28.7	貴	34.0
8	湘	5.5	棠	15.1	椿	22.7	傑	29.0	鐸	34.2
9	楷	6.0	謙	15.6	釗	23.0	榕	29.3	翰	34.5
10	堃	6.6	溥	16.0	均	23.4	坤	29.5	芬	34.7
11	震	7.2	泰	16.4	琳	23.7	增	29.8	藻	34.9
12	杰	7.7	燦	16.8	雲	24.0	澹	30.1	模	35.2
13	銘	8.2	斌	17.2	華	24.3	燾	30.4	忻	35.4
14	彬	8.6	澍	17.6	瑞	24.7	煌	30.6	棟	35.7
15	霖	9.1	熙	18.0	鉞	25.0	琛	30.9	寅	35.9
16	森	9.6	英	18.4	林	25.3	焜	31.2	濟	36.1
17	俊	10.1	煒	18.7	元	25.6	桂	31.4	淳	36.3
18	楨	10.6	瀚	19.1	灝	25.9	蘭	31.7	濂	36.6
19	鼎	11.0	照	19.5	珍	26.3	銑	32.0	塏	36.8
20	銓	11.5	錦	19.9	璜	26.6	麟	32.2	錕	37.0

Note: Based on authors' calculations on 514,417 CGED-Q JSL records with a legible surname and a legible one-character given name.

The given names recorded in the CGED-Q JSL should otherwise be stable and in our experience are the ones recorded for officials in their family genealogies and other sources like the CGED-Q ER, not their courtesy name (*biaozi*, 表字) or style name (*hao*, 號). We have shared data with researchers who have constructed datasets from lineage genealogies, and they report success linking men in the genealogies to officials in the CGED-Q JSL based on the names in the genealogies. Users of our CGED-Q JSL search page also report success locating ancestors or other figures based on names recorded in genealogies or other sources.¹⁷ As for the stability of names, while we have not explicitly searched for cases where an official appeared to change their given name, we are not aware of any cases where someone appeared with two different given names except as the result of problems with the sources or transcription process that we discuss below.¹⁸

3.1.3 PLACE OF ORIGIN

For place of origin, the available level of detail differed between the civil officials recorded in the *Jinshenlu* and the military officials in the *Zhongshubeilan*. The place of origin was where an official had first sat for an exam. In most cases this was where their family lived, but as we will discuss below, there were exceptions. 95% of the records of civil officials with surnames in the *Jinshenlu* specified county of origin and either specified province of origin or allowed for it to be inferred from the province in which the official was currently serving.¹⁹ Of the records of military officers with surnames in the *Zhongshubeilan*, 13% had both province and county of origin, 84% only had province of origin and 3% had county of origin.

17 The search page is located at <http://vis.cse.ust.hk/searchjssl/>.

18 If evidence emerges that officials did change their name, we will have to revisit our procedures for linkage within the CGED-Q JSL to produce career histories, as well as our procedures for linkage to other sources like the CGED-Q ER.

19 Province of origin could be imputed from province of current post because the *Jinshenlu* typically omitted province of origin for officials serving in their home province.

For civil officials, the place of origin was diverse, though not as diverse as the given name. Table 7 presents the cumulative percentages for the 100 most common places of origin as recorded for officials with surnames in the *Jinshenlu*. In most cases this is the province and county or prefecture where an official earned the *shengyuan* (生員) degree that made them eligible to sit for further exams or purchase the degrees that would qualify them for office. Since usually the county was recorded, not the prefecture, below we will only refer to county. A total of 10,156 distinct combinations of province and county or prefecture appeared in the CGED-Q JSL. For reasons that we discuss below, this is larger than the actual number of counties and prefectures at any given time.

The top 10 places of origin accounted for 16.7% of records and the top 100 locations accounted for 45.4% of records. The two most common locations, Daxing (大興) and Wanping (宛平) in Shuntian (順天) require explanation. These were the locations where the sons and possibly other male kin of officials serving in the capital originally sat for their exams. In these and a small number of other cases, the official's family's place of origin was somewhere else.²⁰ As long as province and county of origin of these officials were recorded consistently in every edition in which they appeared as the prefecture in Shuntian where they took the linkage, there is no problem for linkage. Many of the other top counties of origin were in Zhejiang, traditionally an important source of exam passers, degree purchasers, and officials. Other top counties of origin included Changsha (長沙) in Hunan (湖南) in fifth place, Tianjin (天津) in Zhili (直隸) in seventh place, and Chengdu (成都) in Sichuan (四川) in ninth place.

Table 7 Cumulative percentages of the 100 most common places of origin of civil officials with surnames

	1-20		21-40		41-60		61-80		81-100	
	Province and County	%	Province and County	%	Province and County	%	Province and County	%	Province and County	%
1	順天大興	5.2	湖北漢陽	23.4	四川華陽	31.6	山西平定	37.5	江西建昌	41.9
2	順天宛平	7.5	江蘇吳縣	23.9	廣東順德	31.9	四川重慶	37.7	山東歷城	42.1
3	浙江山陰	9.6	湖南善化	24.3	陝西同州	32.2	江西新建	38.0	浙江餘姚	42.3
4	浙江會稽	11.1	貴州貴陽	24.8	直隸河間	32.5	山西介休	38.2	江西南豐	42.5
5	湖南長沙	12.2	山東濟南	25.2	廣東嘉應	32.9	浙江慈谿	38.5	湖南湘潭	42.7
6	浙江仁和	13.2	福建閩縣	25.7	江蘇元和	33.2	安徽合肥	38.7	直隸清苑	42.9
7	直隸天津	14.2	河南開封	26.1	安徽涇縣	33.5	廣東番禺	38.9	陝西長安	43.1
8	浙江錢塘	15.1	陝西西安	26.5	雲南昆明	33.8	直隸永平	39.2	河南固始	43.3
9	四川成都	15.9	廣西桂林	26.9	廣東肇慶	34.1	江蘇金匱	39.4	安徽太平	43.5
10	浙江山陰	16.7	浙江紹興	27.3	安徽歙縣	34.4	安徽甯國	39.6	安徽懷甯	43.7
11	廣東廣州	17.4	浙江蕭山	27.8	山西汾州	34.7	江蘇無錫	39.8	江蘇常州	43.9
12	浙江歸安	18.1	福建侯官	28.2	江蘇長洲	35.0	直隸保定	40.1	江蘇蘇州	44.0
13	安徽桐城	18.8	河南祥符	28.6	河南光州	35.3	江蘇常熟	40.3	浙江秀水	44.2
14	江西南昌	19.5	廣西臨桂	29.0	浙江烏程	35.6	江西南城	40.5	山東武定	44.4
15	福建福州	20.1	浙江杭州	29.4	山西太原	35.9	安徽婺源	40.7	廣東香山	44.6
16	江蘇上元	20.8	浙江嘉興	29.8	廣東南海	36.1	山東諸城	40.9	湖北黃州	44.7
17	江蘇陽湖	21.3	湖北武昌	30.1	江蘇吳縣	36.4	浙江上虞	41.1	河南南陽	44.9
18	江蘇武進	21.9	貴州貴筑	30.5	山東萊州	36.7	江西吉安	41.3	江蘇儀徵	45.1
19	順天通州	22.5	江蘇江甯	30.9	雲南臨安	37.0	直隸順天	41.5	江西新城	45.3
20	湖北江夏	23.0	江蘇丹徒	31.2	山東登州	37.2	貴州遵義	41.7	河南衛輝	45.4

Note: Based on 2,615,955 records of officials with surnames in the *Jinshenlu* with both a province and county of origin recorded. We exclude military officials in the *Zhongshubeilan* because they rarely had county of origin included.

There were two main reasons that the number of combinations of province and county appearing in the data was larger than the number of counties at any given time, and these require attention during

20 There were too few secondary places of origin included to be of much use in linkage. Of the records that included a province and county of origin, only 13,533 (0.39%) listed an additional place of origin.

linkage. First, the province of origin listed for an official could change between editions even when the county of origin did not.²¹ Out of 1,789,985 pairs of records in adjacent editions with identical surname and given name, position, and degree qualification, there were still 0.1% (1941) in which the province changed. This could occur because a provincial boundary was redrawn, but in other cases it was likely the result of a mistake during the production of the edition or the transcription by coders. Several sets of adjacent provinces stood out for the frequency with which one was replaced by the other across two records of the same official with the same county of origin listed: 1) Guangdong and Guangxi, 2) Zhejiang, Jiangsu, Jiangxi, and Anhui, 3) Hubei and Hunan, 4) Shandong and Shanxi, 5) Shuntian and Zhili, and 5) Shaanxi and Gansu.²²

Second, the characters used to write the name of a county could differ across editions. Out of 1,581,616 pairs of records in adjacent editions that had an identical surname, given name, province of origin, degree qualification, and position recorded, there were 3.6% (57,066) in which the county differed. Almost all of these were situations where a character within a county name was replaced with a variant form of the same character, as happened above with the surnames and characters that were part of given names. For example, the third and 10th most common counties (山陰 and 山陰) in Zhejiang (浙江) are the same county (Shanyin) but the second character of the county's name appears in the original source in two different forms. Similarly, the 22nd and 57th most common counties (吳縣 and 吳縣) in Jiangsu (江蘇) are the same county (Wuxian), but the first character appears in the original in two different forms. Other examples include Qiantang (錢塘 and 錢塘) in Zhejiang and Qingyuan (清苑 and 清苑) in Zhili.²³

The cumulative implication of the discrepancies for surname, given name, and location for nominative linkage across the career of all the records of an official across their career is serious. By combining the discrepancy rates for the primary attributes, we can produce estimates that for two records of the same official in two adjacent editions, at least one of the four primary attributes differs. Assuming independence between the probabilities of each of the four primary attributes differing, we have $1-(1-0.035)(1-0.001)(1-0.0434)(1-0.0128) = 0.0896$, or 8.96%. Assuming a typical career length of five years, or 20 quarterly editions, the probability of a discrepancy in at least one pair of records is 83.2% ($1-(1-0.0896)^{20}$). In other words, assuming independence of these probabilities, it is almost certain that for any official whose career lasted for more than a few years that at least one of their records will not match exactly, and in the absence of measures to accommodate discrepancies, the records of many if not most officials with careers of more than just a few years of service will be split incorrectly into two or more officials. Below, we will present tabulations from career histories of officials produced by our linkage to show that such discrepancies were indeed common.

3.1.4 SECONDARY ATTRIBUTES

Secondary attributes help adjudicate in situations where the primary attributes in a pair of records of officials with a surname are close but not an exact match. As we discuss below, they may be useful to confirm a candidate match, but by themselves they are rarely adequate to rule one out because they are not recorded completely, may not be recorded in a consistent fashion, or may change. For example, commercial editions tended to recorded more details that could be used as secondary attributes than

21 The *Zhongshubeilan* editions had additional complications. In the *Zhongshubeilan* rosters of military officials, Huguang (湖廣) appeared as a province of origin in some late 18th century and early 19th century editions. This was a combination of Hunan and Guangdong. We assigned the four counties that were associated with Huguang to Hunan. These were 慈利 (Cili), 祁陽 (Qiyang), 衡陽 (Hengyang), and 道州 (Daozhou). Similarly, in the *Zhongshubeilan* and sometimes in the *Jinshenlu*, counties in Jiangsu, Zhejiang, Anhui and sometimes Jiangxi were listed as being in Jiangnan (江南).

22 When we compared records in adjacent editions that were less than three years apart and which were identical on the surname, given name, county of origin, degree qualification and position, there were 36 cases where an official from Lingui (臨桂) county was listed as being from Guangdong in one record and Guangxi in another, 25 cases where someone was listed with Changping (昌平) as county of origin and were listed as being from Shuntian province in one record and Zhili province in the other, 21 cases where an official from Dantu (丹徒) county was listed as being from Jiangsu in one record and Jiangxi in the other, and 19 cases where an official from Hanyang (漢陽) was listed as being from Hubei in one record and Hunan in the other. Counties that switched between Shuntian and Zhili in more than 10 cases included Baoding (保定), Wuqing (武清), Ninghe (甯河) and Wanping (宛平).

23 We have made a list of pairs of discordant counties available at the same website as the other tables.

official editions (Chen et al., 2020). Available secondary attributes for officials with surnames include the exam or purchased degree that qualified an official for appointment, the official position, courtesy or style name, and title.

The most important of these are degree qualifications. 84.2% of the records of officials with surnames included the examination or purchased degree that qualified them for appointment (*chushen*, 出身). For some officials who held a *jinshi* or *juren* examination degree, the name of the degree wasn't included in the record, but the year (*gan zhi*, 干支) in which they earned their degree was included. Since the provincial and metropolitan exams were the basis of the *juren* and *jinshi* were held in different years, whether an official held a *juren* or *jinshi* could be inferred from the exam year. When *jinshi* or *juren* inferred from exam year are included, 93.2% of records of officials with surnames specified a degree qualification. Hundreds of different degrees were recorded in the original, but for 89.3% of them, the degree fell into one of the following five broad categories: 1) *Jinshi* (進士) degrees for graduates of the Metropolitan Exam, 2) *Juren* (舉人) degrees for graduates of the provincial exam, 3) Regular *gongsheng* (正途貢生) degrees earned by examination, 4) Irregular *gongsheng* (異途貢生) degree acquired by purchase, or 5) Purchased *jiansheng* (監生) degree.²⁴ Of 1,405,138 pairs of records in adjacent editions that matched on surname, given name, place of origin, and post and which had a degree qualification recorded in the original source, only 7.5% (106,007) changed their degree between two editions. Nearly all these changes were within the broad categories above and represented different ways of writing the same degree. Actual transitions between broad categories were rare.²⁵

Official post is useful for confirmation of candidate matches. Relevant information includes an official's job title (*guan zhi*, 官職). For officials in the capital, their ministry and department were recorded. For officials outside the capital, their province, prefecture, and county were recorded. According to our calculations based on record pairs in adjacent editions that were identical on all primary attributes, 7.3% of job titles changed between editions, either because the official changed jobs, or because the title was written differently. If we consider the entire post, including the geographic location or ministry and department, 12.6% changed between editions. Again, this reflected not only actual changes, but inconsistencies across editions in recording. The recorded post had high specificity: for 85% of the records of officials with a surname, the combination of geographic location or ministry and department and job title was unique within the quarterly edition. We have also mapped posts to the numeric bureaucratic ranks used in the civil service (*pin ji*, 品級) and then categorized these numeric ranks as high, middle, low, and unranked. Below, this helps us assess whether two records with the same name belong to the same or different officials.²⁶

Some other attributes were recorded only for a few officials, but when they were recorded, could be useful for helping to confirm a match. One of these was the official's courtesy name (*biao zi*, 表字) or style name (*hao*, 號). 11.7% of the records of officials with a surname included a courtesy or style name alongside the given name. Whether or not these names were recorded also varied across editions: In 74 of 275 *Jinshenlu* editions, no courtesy or style names were recorded at all. They are also not systematically available in the CGED-Q ER, limiting their usefulness for linkage to that dataset. Titles (*ju wei*, 爵位) were recorded consistently, but only 0.5% of civil officials with a surname had one. Year of appointment to the current post and related information could be useful but they are only available for 60.2% of records of officials with a surname in the CGED-Q JSL, and not available at all in the CGED-Q ER. 57 *Jinshenlu* editions do not record year of appointment to the current post.

24 A small number of civil officials in the *Jinshenlu* and many military officials in the *Zhongshubeilan* had military exam (武舉) degrees. A small number of officials were *yinsheng* (蔭生), that is holders of a hereditary honorary status. See Chen et al. (2020) for a detailed discussion of these degrees, including tabulations and trends over time.

25 When transitions between broad categories did occur, they were upward, occurring when an official passed a higher exam while serving. The most common was from *jiansheng* to *juren*, of which there were 1022 cases. There were 633 transitions from *juren* to *jinshi*.

26 When officials held two or more posts at the same time, they tended to be within the same rank category or in adjacent categories. Similarly, when officials changed post between editions, it was usually between posts in the same or adjacent rank categories. Transitions from high to low or high to unranked were extremely rare.

3.2 ATTRIBUTES AVAILABLE FOR OFFICIALS WITHOUT SURNAMENAMES

3.2.1 GIVEN NAMES

26,727 distinct given names appeared for officials without surnames in the data. In principle, all or almost all of these officials should have been Bannermen, mostly Manchu but in some cases Mongol. 84.1% of the given names consisted of only two characters, 11.2% three characters and less than 1% four or more characters. According to Table 8, the top 100 names accounted for 8.6% of records. This was only slightly higher than the 6.6% accounted for by the top 100 given names of officials with a surname. The main difference is that the distribution of given names of officials without surnames has a shorter tail: separate calculations reveal that the top 200 account for 13%, the top 1,000 account for 36%, and the top 10,000 account for 92%. By contrast, the top 10,000 names accounted for only 64% of the records of officials with surnames. While the smaller number of officials without surnames may have accounted for the overall smaller number of distinct given names, it should not have affected the shape of the distribution.

Table 8 *Cumulative percentages of the top 100 most common given names of officials without surnames in the CGED-Q JSL*

	1-20		21-40		41-60		61-80		81-100	
	Given name	%	Given name	%	Given name	%	Given name	%	Given name	%
1	文光	0.2	錫麟	2.4	恩壽	4.1	恒安	5.7	明安	7.0
2	祥麟	0.3	松林	2.5	德興	4.2	恒昌	5.7	慶昌	7.1
3	玉山	0.4	桂森	2.6	祥安	4.3	慶雲	5.8	崇勳	7.1
4	英俊	0.6	瑞麟	2.7	文海	4.4	玉崑	5.9	文溥	7.2
5	文英	0.7	松齡	2.8	延齡	4.5	奎文	5.9	桂斌	7.3
6	文明	0.8	文治	2.9	吉昌	4.5	恩慶	6.0	恩承	7.3
7	長春	0.9	恩光	3.0	崇福	4.6	祥瑞	6.1	定保	7.4
8	慶安	1.0	鍾秀	3.0	恩榮	4.7	祥泰	6.2	清安	7.4
9	慶福	1.2	榮慶	3.1	玉衡	4.8	榮桂	6.2	長慶	7.5
10	毓秀	1.3	常明	3.2	松壽	4.8	文成	6.3	文斌	7.6
11	奎英	1.4	松秀	3.3	文桂	4.9	文惠	6.4	桂昌	7.6
12	恩霖	1.5	文貴	3.4	榮昌	5.0	雙福	6.4	全福	7.7
13	扎拉芬	1.6	慶恩	3.5	榮恩	5.1	佛爾國春	6.5	英奎	7.7
14	英秀	1.7	榮安	3.6	景福	5.1	德馨	6.6	慶祥	7.8
15	慶麟	1.8	崇禧	3.6	景昌	5.2	春慶	6.6	托克托布	7.9
16	德祿	1.9	文瑞	3.7	吉順	5.3	恩明	6.7	英麟	7.9
17	慶瑞	2.0	興奎	3.8	恩隆	5.4	麟祥	6.7	文敬	8.0
18	崇恩	2.1	文麟	3.9	德麟	5.4	桂芬	6.8	常興	8.0
19	桂林	2.2	文秀	4.0	榮光	5.5	德克精額	6.9	松年	8.1
20	文興	2.3	桂芳	4.1	恩綸	5.6	文俊	6.9	全順	8.2

Note: Based on 811,580 records of officials without surnames in the CGED-Q JSL.

The given names recorded for officials without surnames were transliterations into Chinese of originally Manchu or Mongol names. Bannerman officials had different combinations of characters to choose from for the transliteration of their name. For example, the most common name in term of toneless pronunciation, Qing'an, appeared variously as 慶安, 清安, and 清安. In the latter two, 清 and 清 are variants of the same Chinese character. The next most common name in terms of toneless pronunciation, Xilin, appeared as 錫麟, 錫霖, 熙麟 and 西林. These are all different characters. As a result, our tabulations of the romanized names without tones reveals that they were less diverse than names written as Chinese characters. There were 14,560 distinct names if we only consider the pronunciations without tones. The top 100 accounted for 11.8% of records, the top 200 accounted for 19.4% of records, the top 1000 accounted for half of records and the top 10,000 accounted for 99.0% of records.

In the CGED-Q JSL, changes in the transliterations of the same Manchu or Mongol name across different editions appear to have been rare. While officials who had the same Manchu or Mongol name may

have had different transliterations to choose from at the beginning of their career, once they chose one they do not seem to have changed it later. Of 560,559 pairs of records of officials without surnames in editions no more than one year apart that were identical in terms of the toneless Mandarin pronunciation of the name, Banner affiliation, and post, the Chinese characters used to write the name changed in only 2.3% of pairs (13,128). Our further inspection revealed that many of these apparent changes were the result of replacement of one character in the name with a variant form of the same character.

3.2.2 BANNER AFFILIATION

Banner affiliation was stable enough to help confirm candidate links, but there were enough changes to suggest caution against reliance on it to exclude possible links. Every Bannermen were associated with one of eight banners defined by a combination of either Plain or Bordered and one of four colours: Yellow, White, Red, and Blue.²⁷ When we examined 488,734 pairs of records of Manchu and Mongol Bannermen in adjacent editions with identical names in Chinese characters, identical location or ministry and department, and identical job title, 4.4% (21,634) changed banner. More than one-quarter of these were between Plain and Bordered Banners of the same colours. Most of the changes are among officials with the same three job titles as above: clerk (*bitieshi*, 筆帖式), *yuanwailang* (員外郎) or *zhushi* (主事). At present we are unclear of the process by which officials changed Banners, and we will need to conduct further inquiries with the help of Qing historians.

3.2.3 SECONDARY ATTRIBUTES OF BANNERMEN

The posts recorded for officials without surnames within a quarterly edition were not unique. Table 9 presents the tabulation of the concatenation of job title and administrative unit for officials without surnames. For those serving in the capital, the administrative unit was their ministry and department. For those serving outside the capital, it was the province and possibly prefecture and county where they were assigned. Only 16.7% of job titles (*guan zhi*, 官職) were unique within an edition. More than three-quarter appeared five or more times within an edition. The most common were clerks (*bitieshi*, 筆帖式), *yuanwailang* (員外郎) and *zhushi* (主事). Even when we consider the combination of location or ministry and department and job title, less than one-third of positions were unique. For more than half of positions, there were 5 or more records in the same edition with an identical position. Most of the repeated positions were clerks who were in pools assigned to the central government ministries.

Table 9 *Uniqueness of given names and posts for officials without surnames in the Jinshenlu*

Repetitions within edition	Job title (<i>Guanzhi</i> , 官職)	+ Location or Ministry and Department
	%	%
1	16.7	31.1
2	2.5	8.5
3	1.2	4.0
4	1.4	3.8
5	78.2	52.7
Total	100	100
Records	784,502	784,502

Officials without surnames had other details recorded that are potentially useful as secondary attributes, but which are only available for small numbers of records. Those who were members of the main line (*zongshi*, 宗室) or collateral line (*jueluo*, 覺羅) of the Imperial Lineage were recorded as such and accounted for 7.4% of the civil officials who had no surname and 1.7% of civil officials overall. Over the entire course of the Qing and into the Republican era, the Imperial Lineage only had 83,656 male members total, thus its members were heavily overrepresented among officials. One-third (35.8%) of civil officials who were Bannermen had an examination or purchased degree recorded. This tended to be more common later in the 19th century. 11.6% of Bannermen had a courtesy or style name recorded. Year of appointment is only recorded in 7.5% of the records of Bannermen.

27 The upper three were Bordered Yellow (鑲黃旗), Plain Yellow (正藍) and Plain White (正白旗). The lower five were Plain Red (正紅旗), Bordered White (鑲白旗), Bordered Red (鑲紅旗), Plain Blue (正藍旗) and Bordered Blue (鑲藍旗).

4 CHINA GOVERNMENT EMPLOYEE DATASET-QING EXAMINATION RECORDS (CGED-Q ER)

The China Government Employee Dataset-Qing Examination Records (CGED-Q ER) consists of records of examination degree holders transcribed from originally separate lists of exam passers from different sittings of the exam. The most important sources are lists in books self-published by the exam degree holders who had passed at the same sitting of an exam and thought of themselves as classmates. Most of these were titled *Tongnianchilu* (同年齒錄), though some appeared with other titles. Hereafter we refer to them as Classmate Books. Each one listed the surname, given name, and province and county of origin for exam passers at a single sitting along with their current post, if any, and names and degrees held for their father and paternal grandfather and great-grandfather. In most cases they also provide age at passing the exam. They also list other kin, but such information is less systematic. Most of the Classmate Books we have transcribed are for *jinshi* (進士) degree holders who passed the Metropolitan Exam (*Huishi*, 會試) held every three years in the capital and *juren* (舉人) degree holders who passed the Provincial Exam (*Xiangshi*, 鄉試) that qualified them to sit for the Metropolitan Exam. We have also entered similar books for holders of the *Gongsheng* (貢生) degree. For Classmate Books, at present we have entered 5,724 *jinshi* records, 26,870 *juren* records, and 11,990 other records.

We also have less detailed official records of the passers of the Provincial and Metropolitan Exams. For the Provincial Exams, *Xiangshilu* (鄉試錄) rosters record the surname and given name, county of origin, exam rank, and ages of passers of a single sitting. Province of origin is inferred from the location of the exam. Some of these are for sittings of provincial exams for which we also have Classmate Books and are therefore redundant. For all passers of the Metropolitan Exam, the *Jinshi Timinglu* (進士題名錄) lists exam year, surname and given name, province, and county of origin, and ranks in the Metropolitan Exam and the follow-up Court Exam (*Dianshi*, 殿試). For many 19th century sittings of the Metropolitan Exam, we already have Classmate Books that provide more detailed information, thus the *Jinshi Timinglu* is useful mainly for its records of *jinshi* not covered by Classmate Books.

For the CGED-Q ER, we have two linkage tasks. The first is to link records of the same degree holder across Classmate Books and the official records *Xiangshilu* and *Jinshi Timinglu*. This facilitates deduplication of records in situations where we have multiple records of the same exam. This could occur if we have *Xiangshilu* and Classmate Books for the same sitting of a Provincial Exam, or if a sitting of the Provincial Exam is covered by a Classmate Book specific to that sitting and a separately published Classmate Book from the same year that compiles results from multiple provinces. Within the CGED-Q ER, we can also link between the different levels, connecting the records of *juren* to their records as *jinshi*. This allows us to examine how characteristics of a *juren* influenced their chances of going on to earn the *jinshi*. The second task is to link the information about degree holders in the CGED-Q ER to their career records in the CGED-Q JSL. This allows us to examine how the characteristics of degree holders including their family background and their exam performance affected their chances of being appointed subsequently being promoted.

For these linkage tasks we make use of surname, given name, province and county of origin, the year in which the degree was earned, and the type of degree recorded in the CGED-Q JSL and ER. Issues related to the use of surname, given name, and province and county of origin are similar to those in the CGED-Q JSL. The combination of surname, given name, and province and county of origin is almost always unique for degree holders with surnames who earned their degrees at the same time, thus we do not repeat the detailed analysis for officials in the CGED-Q JSL from above. There is also the possibility that across different sources, characters may be replaced by variants. The approach we describe below for dealing with this in the CGED-Q JSL will also work for linkage of exam records. Exam year is useful because it allows us to constrain matching to exclude situations where someone appears to earn the *jinshi* before the *juren*, or else earns it more than a decade after the *juren*.

5 LINKAGE

We carry out linkage in four stages. First, as we describe in 5.1, we prepare for linkage by constructing standardized versions of key attributes. Second, as described in 5.2, we carry out simple deterministic linkage to form groups of records that match exactly on a variety of primary and secondary attributes

and therefore are unambiguously the same official. We then extract the first record in each group to produce the dataset that will be used in the later stages. This substantially reduces the number of records to be considered in the later stages. Third, as described in 5.3, we make use of the capability in the STATA probabilistic linkage package *dtlink* (Kranker, 2018) to specify attributes to be used for 'blocking', according to which pairs of records are selected for scoring in the probabilistic linkage only if they have an exact match on those attributes. By excluding large numbers of record pairs that are clearly not matches, for example ones in which records differ on both surname and given name, it yields another order of magnitude reduction in the time required for linkage. In the fourth stage (5.4), we carry out probabilistic linkage, again with *dtlink*. Candidate pairs of records left over after the formation of record groups and application of blocking are scored and then based on these scores, linked together by assignment of a unique identifier to all records that have been associated with a specific official.

5.1 PREPARATION

We prepare the datasets for linkage by producing standardized versions of the primary and secondary attributes. To reduce the chances that inconsistencies in the recording of a given name for the same person across different editions will produce false negatives during linkage, we create transformed versions of the surname and given name. We begin by consolidating the characters in surnames and given names recognized in the Unicode standard as different versions of the same character.²⁸ Examples in Table 6 include 清 and 淸 (Qing) and 勳 and 勳 (Xun). We refer to these as the CV versions of the names, for 'Consolidated Variants'. We then carry out a second round of consolidation on the CV versions which we group sets of characters in given names that are not recognized as variants in the Unicode standard but look like each other.²⁹ Examples include the ones mentioned in the discussion of Table 6: 傅 (Fu) and 傅 (Chuan), 思 (Si) and 恩 (En), 增 (Zeng) and 曾 (Ceng), and 先 (Xian) and 光 (Guang). We refer to these as the SC versions, for 'Similar Characters'. At the end of the process, each record contains the given name as originally entered, and fields for the CV and the SC version.

We also produce standardized versions of the surnames. We first consolidate variant forms of characters based on the Unicode standard to produce CV versions. We then consolidate similar looking CV characters to produce SC versions. To do this, we manually reviewed the results of the tabulation that produced Table 3 to identify the most common discordant pairs that were not variant forms that would be accounted for by consolidation on the Unicode standard. As we noted in our discussion of Table 3, there were pairs of characters that were different enough that we concluded that they may have been for different people who were otherwise similar on the attributes we matched on. After excluding these, for the time being we have settled on 12 sets of characters that were especially like to appear in place of each other, and which we thought were similar enough that they could be swapped by mistake between editions, either during the production of the editions, or during transcription by our coders.³⁰ This is a more conservative approach than we took with the characters in given names because surnames are less diverse than the characters in given names, and accordingly the risk is higher that two people who are the same on other attributes but differ on their surname really are different people. We may adjust our approach later.

We produce standardized versions of the province and county of origin. We create two versions of the county name romanized by Hanyu pinyin to account for the possibility that characters in the name of a county were replaced with homonyms by mistake. These are listed in Table 10. The first version (PY) includes tone marks, and the second version (PY TL) excludes them. Finally, to address inconsistency in the association of counties with provinces, we create a version of province of origin in which Anhui, Jiangsu, Jiangxi and Zhejiang are all combined into Jiangnan, and Hunan, Guangdong, and Guangxi

28 This includes converting characters mistakenly typed in simplified form into traditional form. See <https://unicode.org/reports/tr38/> for a report on the latest version of the Unicode Han Database. We downloaded the Unicode database for Han Chinese characters from <https://www.unicode.org/Public/UCD/latest/ucd/Unihan.zip>

29 We did this by carrying out a tabulation like the one that produced Table 6 but which only used the CV versions of the characters to produce a list of pairs of characters that are commonly swapped. We manually assessed each of the resulting pairs to flag those that were visually similar enough that it is plausible that they could be switched. We use the resulting pairs to map sets of similar characters to a single character.

30 These were 1) 宋, 朱, 宗, 2) 段, 段, 3) 王, 汪, 江, 4) 馬, 馮, 馮, 5) 柳, 柳, 6) 季, 李, 7) 龍, 龔, 8) 余, 徐, 涂, 9) 湛, 湛, 10) 寇, 寇, 11) 樂, 樂, and 12) 褚, 褚.

are all combined into Huguang. We refer to this as the C version of province. In the very small number of records in which a second province and county of origin were listed in the original source, we used that instead of the first listed province and county of origin.

5.2 DETERMINISTIC LINKAGE

We group records that match exactly on a large number of primary and secondary attributes and are in editions less than one year apart and create an extract of the data that only includes the first record in each of these groups. We make the criteria for inclusion of a record in one of these groups so exacting as to rule out false positives in which records of different officials are accidentally linked.³¹ The creation of these record groups by deterministic linkage is straightforward and we do not discuss it further. Because the number of record groups that need to be linked is an order of magnitude less than the original number of records to be linked, the time required for the second and third stages is substantially reduced.

5.3 BLOCKING

We divide blocking for the CGED-Q JSL and CGED-Q ER linkage into six types based on the attributes available in the records involved and the risks of false positives or negatives. For linkage within the CGED-Q JSL to produce career histories, we distinguish three types: 1) officials with a surname who had a single character given name, 2) officials with a surname who had two character given names, and 3) officials without surnames. We link officials with surnames and one-character given names separately because comparison of Tables 4 and 5 suggests that the risk of a false positive is higher, compared with the ones with a two-character given name. This requires stricter criteria for matching on other attributes. Because the combination of surname and two-character given name is more likely to be unique, for linkage of officials with given names who had two-character given names we can be more forgiving for other attributes. Officials without surnames have only the given name and Banner affiliation as primary attributes, which combination is less likely to be unique, thus we must put more weight on secondary attributes. Linkage within the CGED-Q ER forms the fourth type. Here, we treat all the records the same. The total number of records is small enough that false positives for degree-holders with a surname and only a one-character given name are unlikely. For linkage between the CGED-Q JSL and CGED-Q ER, we distinguish the fifth and sixth types according to whether men with surnames have one- or two-character given names.

Table 10 summarizes the attributes used for blocking for each of the six types of linkage. In each case, we balance the risk of false negatives associated with use of overly strict criteria against the increased linkage time associated with the use of loose criteria. In general, we make the blocking criteria as loose as possible while seeking to prevent clearly impossible pairs through to be scored. Thus, for example, we typically block on SC versions of names rather than CV versions of names, and then use scoring on other attributes to assess pairs that match on the SC but not CV versions. For blocking within the CGED-Q JSL, we apply different criteria for each linkage type. For the first type, officials with surnames who had two-character names, we block on the SC and pinyin versions of the surname and given name. That is, if two records have the same SC or pinyin version of the surname and given name, they are a candidate match and go on to be scored on the other attributes, including the CV versions of the names. We do not use the CV version of the names for blocking because it would be too strict, and would preclude making matches based on the looser criteria associated with use of the SC versions. For the second type, officials with surnames and only one-character given names, we only allow pairs of records with the same SC versions of the names. Our experiments with allowing for matches on the pinyin version of the surname and given name yielded too many false positives. For the third type, officials with no surname, we block on the SC version of the given name and Banner affiliation, or on the combination of the pinyin version of the name, the Banner affiliation, Imperial Lineage affiliation, title, and complete post. In other words, a pair in which the SC version of the name doesn't match but the pinyin version does match can still be treated as a candidate pair and scored if there is an exact match on a variety of other characteristics. We allow candidate pairs that match on the pinyin name

31 For officials with surnames, records in a group must have the same CV version of the surname and given name, the same C version of the province of origin, and the same pinyin for the county of origin. For Bannermen, records in a group must have the same CV version of the given name, the same Banner affiliation, and the same government post, which is the concatenation of the administrative unit and job title. We require records in Bannermen sets to match on post as well because as Table 1 showed, the combination of given name and Banner affiliation was not unique within an edition.

only when several additional secondary attributes also match because allowing candidate pairs based on pinyin given name alone would substantially expand the number of pairs to be considered. For the fourth type, linkage within the CGED-ER, the SC versions of the surname and given name are sufficient for blocking. Rather than have a separate approach to blocking in the CGED-Q ER for men with a surname and a single character given name, as we describe below, we apply tighter criteria for scoring candidate pairs involving such records.³² For the fifth type, linkage between the CGED-Q ER and CGED-Q JSL of men with a surname and a two-character given name, we allow for candidates pairs that match on the SC or toneless pinyin versions of the surname and name.³³ For the sixth type, linkage between the CGED-Q ER and CGED-Q JSL of men with single-character given names between, we only allow candidate pairs that match on the SC versions of the names.

Table 10 *Attributes used for blocking in linkage of the CGED-Q JSL and CGED-Q ER*

Linkage Type	Blocking
Within the CGED-Q JSL	
1 Officials with surnames and two-character given names	Surname SC + Given name SC OR Surname PY+ Given name PY
2 Officials with surnames and one-character given names	Surname SC+ Given name SC
3 Officials without surnames	Given name SC + Banner affiliation OR Given name PY + Banner affiliation + Imperial Lineage status + Noble title + Post
Within the CGED-Q ER	
4 All records	Surname SC + Given name SC
Between the CGED-Q ER and CGED-Q JSL	
5 Men with surnames and two-character given names, and men without surnames	Surname SC + Given name SC OR Surname PY NT + Given name PY NT
6 Men with surnames and one-character given names	Surname SC + Given name SC

5.4 PROBABILISTIC LINKAGE

Since probabilistic linkage is already widely used and described in detail elsewhere, here we only provide a summary of the basic concept. Probabilistic matching considers every possible pair of records in a dataset left over after blocking and then scores each pair for similarity according to criteria specified by the user. For the scoring, the user specifies the attributes to compare, and the amount to be added to or subtracted from the score if they match or differ. Calipers may also be specified according to which some amount may be added to the score for a pair if two numeric attributes are within some range of each other, and some other amount may be deducted if they are not. A match is made by comparing the scores of candidate pairs and selecting the ones with the highest score that also meet a cutoff score set by the user.

We scored the candidate pairs of record groups left over after blocking according to their concordance or discordance on specified primary and secondary attributes. Tables 11 and 12 summarize our current rewards and penalties for concordance or discordance on each primary or secondary attribute for our six types of linkage. The rewards ("+" in the tables) are added to the score for a candidate pair if the condition specified in the row heading is satisfied. The penalties ("- " in the tables) are subtracted from the score if the condition is not satisfied. Tables 11 and 12 also include the cutoffs that a score had to be greater than or equal to in order for a match to be made. For each of the six linkage tasks, we choose the amounts to be added to or subtracted for a match or mismatch on a specified attribute to balance the risks of false negatives and false positives. We apply more stringent criteria when there are larger numbers of records to be linked, most notably within the CGED-Q JSL, and therefore a higher chance that separate individuals will have the same primary attributes. We apply looser criteria when the chances of a false positive are lower, usually because there are fewer records to be linked. Linkage within the CGED-Q ER is one example.

- 32 We do not have separate blocking for Bannermen when linking between the CGED-Q JSL and CGED-ER because there are too few of them (1.2% of records overall) in the CGED-Q ER to warrant special handling.
- 33 We include Bannermen with officials with surnames because only a small proportion (1.23%) of exam degree holders in the Classmate Books we have coded were Bannermen, and the chances of different individuals having the same name were small.

Table 11 *Rewards and penalties for concordance or discordance on attributes in candidate pairs for linkage within the CGED- Q JSL*

	CGED-Q JSL					
	Type 1 Surname and two- character given name		Type 2 Surname and one- character given name		Type 3 No surname	
	+	-	+	-	+	-
Primary attribute						
Surname (CV) + Given name (CV) + County (PY)						
Surname (SC) + Given name (SC) + County (SC)						
Surname (SC) + Given name (SC) + Province (C)						
Surname (CV) + Given name (CV)	100	0	100	0		
Surname (CV) + Given name (SC)						
Surname (SC) + Given name (PY)						
Given name (CV)					50	0
Given name (SC)					50	0
Province (C)	100	-400	100	-400		
County 1 (Original)						
County 1 (PY)	200	-100	200	-100		
County 1 (PY) in Jiangnan, Huguang	0	-200	0	-200		
Banner affiliation					50	-100
Secondary attribute						
Courtesy or Style name	300	0	300	0	200	0
Imperial Lineage status					100	0
Title					100	0
<i>Post</i>						
Province	25	0	25	0		
Ministry, Agency, or Prefecture	25	0	25	0		
Department or County	25	0	25	0		
Job title	25	0	25	0		
Complete	100	0	100	0		
<i>Rank (Pinji) category</i>						
Same					0	-25
Differ by less than 2					0	-50
Differ by less than 3					0	-400
<i>Degree</i>						
Original	50	0	50	0	50	0
Broad category	0	-100	0	-100	0	-100
Broad category in Jiangnan or Huguang	0	-100	0	-100		
<i>Year</i>						
Same					50	0
< 5 years apart					0	-50
< 10 years apart		-100		-100	0	-100
< 20 years apart		-200		-200		
< 30 years apart						
< 40 years apart		-500		-500	0	-400
Surname and Given name of record above	50	0	50	0	50	0
Surname and Given name of record below	50	0	50	0	50	0
Cutoff		100		100		150

Table 12 *Rewards and penalties for concordance or discordance on attributes in candidate pairs for linkage within the CGED- Q ER and between the CGED-Q ER and CGED-Q JSL*

	CGED-Q ER		CGED-Q JSL to CGED-Q ER			
	Type 4		Type 5		Type 6	
	All		Surname and two-character given name, OR no surname		Surname and one-character given name	
	+	-	+	-	+	-
Primary attribute						
Surname (CV) + Given name (CV) + County (PY)			500	0	500	0
Surname (SC) + Given name (SC) + County (SC)	300	0	200	0	200	0
Surname (SC) + Given name (SC) + Province (C)	100	0	200	0		
Surname (CV) + Given name (CV)			150	0	200	0
Surname (CV) + Given name (SC)			100	0	150	0
Surname (SC) + Given name (PY)			50	0		
Given name (CV)						
Given name (SC)						
Province (C)		-200	0	-200		
County 1 (Original)			100	0		
County 1 (PY)	0	-200	100	0		
Year						
Same						
< 5 years apart						
< 10 years apart	0	-100	100	0	100	0
< 20 years apart	0	-300				
< 30 years apart						
< 40 years apart			0	-500	0	-500
Cutoff		100		200		200

We arrived at the rewards, penalties, and cutoffs in Tables 11 and 12 iteratively. We inspected the results every time we ran the linkage. We located false negatives by searching the data for groups of records that matched exactly on secondary attributes such as position and degree and most but not all of the primary attributes, and which were not associated with a single official. We examined these groups to assess whether the records in the group should all have been assigned to the same official. This helped clarify how often characters were replaced with ones that looked similar and inspired our effort not only to create the CV and SC versions of names. It led us to increase the rewards for exact matches on such secondary attributes as courtesy name and complete post that were highly unlikely to match by chance. It also led to our discovery of inconsistencies in the recording of province.

We searched for false positives by identifying groups of records that had all been assigned to the same official, but which differed on at least one primary attribute, for example, surname, or one character in a two-character given name. This led to our realization that we needed to apply more stringent criteria for individuals with single character given names and led us also to increase penalties for mismatches on attributes such as province of origin or broad category of purchased or examination degree that should be stable. Users working with extracts of the data to study topics of their own, most commonly the appointment and promotion of specific categories of officials, also reported problems that they noticed, and our investigations revealed.³⁴

34 For example, the analyses that underpinned Chen et al. (2018), Hu et al. (2020), Hu, Hu, Chen and Campbell (2021) and Xue and Campbell (2022) all led to discovery of problems that were addressed by refinements to linkage procedures.

For linkage within the CGED-Q JSL (Types 1 through 3), we assigned the largest rewards to concordance on attributes like given name, post, or county that are the most diverse and therefore the least likely to match by chance. Even though blocking differed for one- and two-character names, scoring was the same. We gave large rewards to matches on secondary attributes like courtesy or style name and complete post. This helped counter the effects of inconsistencies in the recording of province and county of origin that were not addressed by the transformations described above. Since posts were listed in the same order from one edition to the next, we also rewarded concordance on the name of the official in the record above or below. Rewards are smaller for concordance on attributes like province or broad category of examination or purchase degree that are less diverse and more likely to match by chance.

We apply the largest penalties for discordance on attributes like province or county of origin that should have been stable and were less diverse. A mismatch on a less diverse attribute like the C version of the province, Banner affiliation, or broad category of degree qualification will lead to a large penalty. We apply a penalty for a mismatch on county, with an additional penalty if the province in which the counties are located are part of Huguang or Jiangnan.³⁵ We also penalize matches of records that are further apart in time, and in the case of records so far apart that it is implausible for them to be the same person, we apply a penalty so large that it will preclude a match from being made. For officials without surnames, we also penalize candidate matches if the categories of bureaucratic rank (*pinji*, 品级) are too far apart. This helps reduce the chances that a record of a high official will be linked to those of another officials with the same given name who is a low-ranking clerk. We apply a smaller penalty for mismatches on attributes that are more prone to inconsistent recording, like detailed examination or purchase degree. Courtesy and style names were diverse, often missing, and sometimes seem to have changed, thus we do not apply a penalty for a mismatch on them. Similarly, because complete positions and the components that made up the position were expected to change when an official was promoted or reassigned, and because different editions could record positions differently even when the official was not promoted or reassigned, we do not apply a penalty for a mismatch on position.

For linkage within the CGED-Q ER (Type 4), we began with surname, name, province and county of origin, and exam year. We created CV and then SC versions of the surname and name. We blocked on the SC version of the surname and name. We rewarded matches on the combination of SC surname, SC name, and county or province, and heavily penalized discordance on the pinyin (PY) version of the county or C version of the province. We used the SC version of the name rather than the CV version because the overall number of men to be linked was much smaller than in the CGED-Q JSL and the risk of a false positive accordingly smaller. We applied only a mild penalty for a gap between exam years because we wanted to allow for links between records of *juven* and *jinshi* degrees earned in different years but applied a much larger penalty if the exam years were so far apart that the rules would not have allowed a *juven* to sit for the Metropolitan exam in the specified year.

For linkage between the CGED-Q JSL and CGED-Q ER (Types 5 and 6), we relied on surname, given name, province and county of origin, CGED-Q ER exam year, and CGED-Q JSL edition year. We blocked on the SC version of the surname and given name. We gave very large rewards for matches on the CV version of the surname and name and smaller rewards for matches on the SC versions. We allowed for matches not only on the province and county of origin in the CGED-Q JSL, but also on the province and county of origin (籍贯) listed in the CGED-Q JSL for officials who sat for the exam someplace other than their actual place of origin, usually Shuntian. We allowed up to 30 years for the time between earning a degree and being appointed for the first time.

5.5 RESULTS

To illustrate how the approach describe above reduces false positives and false negatives, while also reducing the amount of time required, we present the results of linkage within the CGED-Q JSL, that is Types 1, 2 and 3. We focus on linkage with the CGED-Q JSL because it was the most challenging and complex, and made use not only of primary attributes, but a wide range of secondary attributes. Linkage of the 4,108,586 records in the CGED-Q JSL with a name and other information required by the approach described in the sections above yielded 326,315 sets of linked records, each a career history of a single official. For each of the three types of linkage within the CGED-Q JSL, Table

35 Because place of origin could change because of the redefinition of administrative units, we set the penalty for mismatch on county so that it can still be offset by rewards for matches on other attributes. A mismatch on county, in other words, doesn't preclude a match if other attributes are in correspondence.

13 presents the original number of records to be linked, the number of groups remaining after the deterministic linkage described in 5.2, the number of candidate pairs left after the blocking described in 5.3, and the final number of officials produced by the probabilistic linkage described in 5.4. According to Table 13, grouping records with deterministic linkage on the primary and some secondary attributes substantially reduces the number of items to be linked. In the case of Type 1 linkage, the number of items to be linked is reduced by 88.6% percent, from 2,676,108 to 315,015. The resulting number of candidate pairs to be scored is modest. For Type 1 linkage, the number of candidate pairs is lower than the number of groups because many groups are isolates: blocking left them without any other groups to be paired with and scored, and they went straight to being recognized as an official. The number of candidate pairs for Type 3 linkage is much larger because only the given name and Banner affiliation are available for blocking, and these are less diverse than the surname, given name and province and county of origin of officials who have surnames.

Table 13 Results for Type 1, 2, and 3 linkage, CGED-Q JSL dataset

	Type 1 Surname and two- character given name	Type 2 Surname and one- character given name	Type 3 No surname
Records for linkage	2,767,108	527,570	813,908
Number of groups after deterministic linkage	315,015	76,885	171,449
Candidate pairs after blocking	199,263	46,231	398,353
Career histories after linkage	218,946	45,965	64,940

Probabilistic linkage on standardized primary attributes that compensates for discrepancies when there are matches on secondary attributes reduces the number of false negatives. Had we required exact matches on the primary attributes as originally recorded, each distinct combination within one of the histories produced by our linkage would have been associated with a separate official. Table 14 tabulates the career histories according to the numbers of distinct combinations of surname, name and province and county of origin or Banner affiliation within them in the original data. In 28% (100-28) of the career histories of officials with a one-character given name, more than one surname, given name, or place of origin appeared. The corresponding figure for officials with two-character given names was 29.9% (100-70.1). In the career histories of officials of without surnames produced by linkage, 13.9% (100-86.1) had more than one name or Banner affiliation appeared. According to our calculations, linkage by requiring exact matching on the original primary attributes and not using probabilistic linkage with the standardized versions of the names consolidated CV or SC versions of names to allow for discrepancies would have led to the creation of 453,375 career histories. Career histories that in our probabilistic linkage were attributed to a single official would have been separated. The total number of career histories, in other words, would have been inflated by 38%. The gains associated with applying probabilistic linkage within the CGED-Q ER and between the CGED-Q JSL and CGED-Q ER are similar: the number of *juren* degree holders who are linked to *jinshi* records increases substantially, as do the numbers of *juren* and *jinshi* linked to the CGED-Q JSL.

Table 14 Combinations of surname, given name, and province and county of origin or banner affiliation in original data within sets of records for officials produced by linkage in the CGED-Q JSL

Distinct combinations of primary attributes observed within career histories produced by linkage	Type 1 Surname and two- character name %	Type 2 Surnames and one- character name %	Type 3 No surname %	Total %
1	70.1	72.0	86.1	73.5
2	20.9	20.0	11.7	19.0
3	5.7	5.0	1.7	4.8
4	2.1	1.8	0.4	1.7
5	1.3	1.4	0.1	1.1
Total	100	100	100	100
Number of officials	218,946	45,965	64,940	329,851

6 CONCLUSIONS

This is unlikely to be the final word, especially for the linkage of officials without surnames. Based on manual examination of the resulting data we are confident that our linkage of officials with surnames is close to optimal in terms of its balance between avoiding false positives and false negatives, and that any further accommodation of additional discrepancies we have noticed would open the door to false positives in which the records of clearly different officials would be combined. Any further adjustments to the linkage of officials with surnames are likely to consist of small refinements to the lists of similar characters, and adjustments to the handling of problems with provinces and counties. For Bannermen, however, we suspect that the lack of diversity in the combination of names and Banner affiliation means that we still have too many false positives.

Our experiences, and our descriptive results about patterns in names, should be useful to other teams that are carrying out large-scale record linkage in datasets constructed from historical Chinese sources. The issues we discuss here and our approach to linkage are most relevant for the linkage of highly structured data transcribed from rosters and related records, the descriptive results on the consistency and potential for overlap in the recording of names may be of interest to those conducting disambiguation in unstructured data like newspaper articles. Particular attention needs to be paid to the possibility that across difference sources, the characters in the names of individuals to be linked may be replaced with variant forms of the same character, or entirely different characters that are superficially similar.

We now have ongoing projects to construct, link, and analyze datasets of individuals during the Republican era (1911–1949). Our efforts to create datasets from university student records are the furthest along (Ren et al., 2020), but we have other projects to create datasets of Republican officials, professionals, and other elites. While we expect some of our experiences with Qing records to be relevant, we also anticipate that there will be other issues specific to the Republican data. Naming patterns may have changed. Consistency in the usage of genealogical given names as opposed to courtesy or style names may have changed as well. Customs for the recording of place of origin may also have evolved.

ACKNOWLEDGMENTS

This research was supported by Hong Kong Research Grants Council General Research Fund 16602621 (Campbell PI). We are grateful to members of the Lee-Campbell Group, especially Hao Dong, Lawrence Zhang, James Lee, and Matthew Noellert, for their feedback and suggestions. We are also grateful to Loretta Kim for sharing her knowledge of Manchu naming practices. We are also grateful to Xue Qin, Chen Jun, and other users of the data who brought issues they discovered to our attention, leading directly or indirectly to adjustments in our linkage procedures.

REFERENCES

- Abramitzky, R., Mill, R., & Pérez, S. (2020). Linking individuals across historical sources: A fully automated approach. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(2), 94–111. doi: [10.1080/01615440.2018.1543034](https://doi.org/10.1080/01615440.2018.1543034)
- Akgün, Ö., Dearle, A., Kirby, G., Garrett, E., Dalton, T., Christen, P., Dibben C., & Williamson, L. (2020). Linking Scottish vital event records using family groups. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(2), 130–146. doi: [10.1080/01615440.2019.1571466](https://doi.org/10.1080/01615440.2019.1571466)
- Armand, C., Guo, W., Henriot, C., Hu, Y., & Van den Bosch, N. (2022). *Modern China Biographical Database (MCBD). User manual*. Aix en Provence: ENP-China, Aix-Marseille University. Retrieved from https://bookdown.enpchina.eu/mcbd_usermanual/
- Bao, H., Cai, H., Jing, Y., & Wang, J. (2021). Novel evidence for the increasing prevalence of unique names in China: A reply to Ogihara. *Frontiers in Psychology*, 12(731244), 1–6. doi: [10.3389/fpsyg.2021.731244](https://doi.org/10.3389/fpsyg.2021.731244)

- Cai, H., Xi, Z., Yi, F., Liu, Y. & Jing, Y. (2018). Increasing need for uniqueness in contemporary China: Empirical evidence. *Frontiers in Psychology*, 9(554), 1–7. doi: [10.3389/fpsyg.2018.00554](https://doi.org/10.3389/fpsyg.2018.00554)
- Campbell, C. D. (2020). Qingmo keju tingfei dui shiren wenguan qunti de yingxiang-jiyu weiguan dashuju de hongguan xin shijiao [The influence of the abolition of the examinations at the end of the Qing on the holders of exam degrees]. *Shehui kexue jikan [Social Science Journal]*, 4(249), 156–166.
- Campbell, C. D., Chen, B., Ren, Y., & Lee, J. Z. (2019). China Government Employee Database-Qing (CGED-Q) Jinshenlu public release [Database]. DataSpace@HKUST, V14. doi: [10.14711/dataset/E9GKRS](https://doi.org/10.14711/dataset/E9GKRS)
- Campbell, C. D., & Lee, J. Z. (2020). Historical Chinese microdata. 40 years of dataset construction by the Lee-Campbell research group. *Historical Life Course Studies*, 9, 130–157. doi: [10.51964/hlcs9303](https://doi.org/10.51964/hlcs9303)
- Campbell, C. D., Lee, J. Z., & Elliott, M. (2002). Identity construction and reconstruction: Naming and Manchu ethnicity in Northeast China, 1749–1909. *Historical Methods*, 35(3), 101–116. doi: [10.1080/01615440209601201](https://doi.org/10.1080/01615440209601201)
- Chen, B. (2019). *Origins and career patterns of the Qing government officials (1850–1912): Evidence from the China Government Employee Dataset-Qing (CGED-Q)* (PhD dissertation). Hong Kong University of Science and Technology Division of Social Science, China.
- Chen, B., & Campbell, C. D. (2023). Cong yizhong dao duozhong shiliao: Lijie Qingdai guanyuan shitu de xin fangfa [From one source to many sources: New methods for understanding the careers of Qing officials]. *Shixue Yuekan [History Monthly]*. Forthcoming publication.
- Chen, B., Campbell, C. D., & Lee, J. Z. (2018). Qingmo xinzheng qianhou qiren yu zongshi guanyuan de guanzhi bianhua chutan-yi jinshenlu shujukou wei cailiao de fenxi [The transition of banner and imperial lineage officials during the late Qing reform period: Evidence from the Qing Jinshenlu Database]. *Qingshi Yanjiu [Studies in Qing History]*, 2018(4), 10–20. Retrieved from <http://qsyj.iqh.net.cn/CN/abstract/abstract2384.shtml>
- Chen, B., Campbell, C. D., Ren, Y., & Lee, J. Z. (2020). Big data for the study of Qing officialdom: The China Government Employee Database-Qing (CGED-Q). *The Journal of Chinese History*, 4(2), 431–460. doi: [10.1017/jch.2020.15](https://doi.org/10.1017/jch.2020.15)
- Chen, M., Du, Q., Shao, Y., & Long, H. (2018). Jiyu yinxingma de hanzi xiangsidu bidui suanfa [Chinese characters similarity comparison algorithm based on phonetic code and shape code]. *Xinxu Jishu [Information Technology]*, 11, 73–75. doi: [10.13274/j.cnki.hdzj.2018.11.016](https://doi.org/10.13274/j.cnki.hdzj.2018.11.016)
- Chen, S., & Wang, H. (2022). China Biographical Database (CBDB): A relational database for prosopographical research of pre-modern China. *Journal of Open Humanities Data*, 8(4). doi: [10.5334/johd.68](https://doi.org/10.5334/johd.68)
- Chen, Y., & Huang, C. (2010). Exploring personal name disambiguation from name understanding. *2010 4th International Universal Communication Symposium*, 345–349, doi: [10.1109/IUCS.2010.5666185](https://doi.org/10.1109/IUCS.2010.5666185)
- Chua, I. (2021). What can we tell from the evolution of Han Chinese names? *Kontinentalist*. Retrieved from <https://kontinentalist.com/stories/a-cultural-history-of-han-chinese-names-for-girls-and-boys-in-china>
- Elliott, M. C., Campbell, C. D., & Lee, J. Z. (2016). A demographic estimate of the population of the Qing banners. *Études Chinoises*, 35(1), 9–40.
- Fan, C., & Li, Y. (2021). Chinese personal name disambiguation based on clustering. *Wireless Communications and Mobile Computing*, 3790176. doi: [10.1155/2021/3790176](https://doi.org/10.1155/2021/3790176)
- Fuller, M. A. (2021). *The China Biographical Database user's guide. Revised version 3.3*. Retrieved from <https://projects.iq.harvard.edu/cbdb/supporting-documents>
- Han, H., Yao, C., Fu, Y., Yu, Y., Zhang, Y., & Xu, S. (2017). Semantic fingerprints-based author name disambiguation in Chinese documents. *Scientometrics*, 111, 1879–1896. doi: [10.1007/s11192-017-2338-6](https://doi.org/10.1007/s11192-017-2338-6)
- Han, W., Xu, X., & Zhao, T. (2011). Study on Chinese person name disambiguation based on multi-stage strategy. *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 1177–1181. doi: [10.1109/FSKD.2011.6019646](https://doi.org/10.1109/FSKD.2011.6019646)
- Helgertz, J., Price, J., Wellington, J., Thompson, K. J., Ruggles, S., & Fitch, C. A. (2022). A new strategy for linking U.S. historical censuses: A case study for the IPUMS multigenerational longitudinal panel. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 55(1), 12–29. doi: [10.1080/01615440.2021.1985027](https://doi.org/10.1080/01615440.2021.1985027)

- Hu, C., Hu, H., Chen, B., & Campbell, C. D. (2021). Qingdai zhou de zhengqu fendeng yu zhizhou xuanren de lianghua fenxi [Quantitative analysis on the local government administrative categorization system and the appointment of department prefects during the Qing]. *Shuzi Renwen Yanjiu [Digital Humanities Research]*, 1(1), 34–47.
- Hu, H., Chen, C., & Campbell, C. D. (2020). Qingdai zhifu xuanren de kongjian yu lianghua fenxi-yi zhengqu fendeng, <jinshenlu> shujuku wei zhongxin [The appointment of prefects during the Qing: A spatial and quantitative analysis focusing on the system of administrative division and using the CGED-Q]. *Xinya Xuebao [New Asia Journal]*, 37, 339–398.
- Kim J., Kim, J., & Kim, J. (2021). Effect of Chinese characters on machine learning for Chinese author name disambiguation: A counterfactual evaluation. *Journal of Information Science, OnlineFirst*. doi: [10.1177%2F01655515211018171](https://doi.org/10.1177/2F01655515211018171)
- Kranker, K. (2018). DTALINK: Stata module to implement probabilistic record linkage. Statistical Software Components S458504, Boston College Department of Economics, revised 16 Feb. 2019. Retrieved from <https://ideas.repec.org/c/boc/bocode/s458504.html>
- Lee, J. Z., & Campbell, C. D. (1997). *Fate and fortune in rural China. Social organization and population behaviour in Liaoning, 1774–1873*. Cambridge: Cambridge University Press.
- Lee, J. Z., Campbell, C. D., & Chen, S. (2010). *China Multi-Generational Panel Dataset, Liaoning (CMGPD-LN) 1749–1909. User guide*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.
- Liu, M., Rus, V., Liao, Q., & Liu, L. (2017). Encoding and ranking similar Chinese characters. *Journal of Information Science and Engineering*, 33(5), 1195–1211. doi: [10.6688%2fjise.2017.33.5.6](https://doi.org/10.6688/2fjise.2017.33.5.6)
- Ren, B., Chen, L., & Lee, J. Z. (2020). Meritocracy and the making of the Chinese academe, 1912–1952. *The China Quarterly*, 244, 942–968. doi: [10.1017/S0305741020001289](https://doi.org/10.1017/S0305741020001289)
- Ren, Y., Chen, B., Hao, X., Campbell, C. D., & Lee, J. Z. (2016). Qingdai jinshenlu lianghua shujuku yu guangliao qunti yanjiu [The Qing Jinshenlu database: A new source for the study of Qing officials]. *Qingshi Yanjiu [Qing History Research]*, 2016(4), 61–77.
- Ren, Y., Chen, B., Hao, X., Campbell, C. D., & Lee, J. Z. (2019). *Zhongguo lishi guanyuan lianghua shujuku-qingdai jinshenlu (1900–1912). Shidian gongkaiban yonghu zhinan. [The China Government Employee Database-Qing (CGED-Q) Jinshenlu (JSL) 1900–1912. Public release user guide]*. doi: [10.14711/dataset/E9GKRS](https://doi.org/10.14711/dataset/E9GKRS)
- Ruggles, S. (2002). Linking historical censuses: A new approach. *History and Computing*, 14(1–2), 213–224. doi: [10.3366/hac.2002.14.1-2.213](https://doi.org/10.3366/hac.2002.14.1-2.213)
- Stone, L. (1971). Prosopography. *Daedalus*, 100, 46–79.
- Sylvester, K., & Hacker, J. D. (2020). Introduction to special issues on historical record linking. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(2), 77–79. doi: [10.1080/01615440.2020.1707445](https://doi.org/10.1080/01615440.2020.1707445)
- Tsui, L. H., & Wang, H. (2020). Harvesting big biographical data for Chinese history: The China Biographical Database (CBDB). *Journal of Chinese History*, 4(2), 505–511. doi: [10.1017/jch.2020.21](https://doi.org/10.1017/jch.2020.21)
- Wang, H., Chen, S., Dong, H., Noellert, M., Campbell, C. D., & Lee, J. Z. (2013). *China multi-generational panel dataset, Shuangcheng (CMGPD-SC) 1866–1914. User guide*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research. doi: [10.3886/ICPSR35292.v9](https://doi.org/10.3886/ICPSR35292.v9)
- Wang, Y., Liang, H., Shu, X., Wang, J., Xu, K., Deng, Z., Campbell, C. D., Chen, B., Wu, Y., & Qu, H. (2021). Interactive visual exploration of longitudinal historical career mobility data. *IEEE Transactions on Visualization and Computer Graphics*, 28(10), 3441–3455. doi: [10.1109/TVCG.2021.3067200](https://doi.org/10.1109/TVCG.2021.3067200)
- Xu, S., Zheng, M. & Li, X. (2020). String comparators for Chinese-characters-based record linkages. *IEEE Access*, 9, 3735–3743. doi: [10.1109/ACCESS.2020.3047927](https://doi.org/10.1109/ACCESS.2020.3047927)
- Xue, Q., & Campbell, C. D. (2022). Qingji gaige shiyu xia libu guanyuan qunti de renshi dishan yu jiegou bianqian (1898–1911) — Yi "jinshenlu" shujuku wei zhongxin [Change and constancy: The personnel evolution and structural change of the ministry of personnel during the reform in Qing dynasty — Based on China Government Employee Database-Qing (CGED-Q)]. *Shehui Kexue Yanjiu [Social Science Research]*, 2(259), 173–182.
- Yin, D., Motohashi, K., & Dang, J. (2020). Large-scale name disambiguation of Chinese patent inventors (1985–2016). *Scientometrics*, 122, 765–790. doi: [10.1007/s11192-019-03310-w](https://doi.org/10.1007/s11192-019-03310-w)

HISTORICAL LIFE COURSE STUDIES
VOLUME 11 (2021), published 10-08-2021

Building Longitudinal Datasets From Diverse Historical Data in Australia

Janet McCalman

Melbourne School of Population and Global Health, University of Melbourne

ABSTRACT

Australia is rich in population datasets generated to manage convicts, civilians, stock, land and the colonised and displaced First Nations people. It has also preserved all service and pension data from both world wars. Through nominal linkage using volunteers and paid research staff, it has been possible over the past twenty years to build four cradle-to-grave datasets derived from administrative cohorts: poor white babies born in a charity hospital 1858–1900; Aboriginal Victorians from 1855 to 1988; convicts transported to Van Diemen's Land 1818–1853 and servicemen who embarked for World War I from the State of Victoria. The abundance of digitised historical sources from government archives to historical newspapers enables the practice of demographic prosopography, with a wide range of variables that have yielded new insights into Australia's population and social history.

Keywords: Prosopography, Life course, Insults, Early life effects, Race, Class

DOI article: <https://doi.org/10.51964/hlcs10939>

© 2021, McCalman

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

Australia is a deeply bureaucratic society. Its origins in separate British colonies, some of them penal settlements, established a culture of accounting of people, things and stock. From the first settlement in 1788, convicts, soldiers and naval personnel had to be managed bureaucratically, and the convict records are on the UNESCO World Heritage Register. The best records have survived in Tasmania (originally Van Diemen's Land) where they amounted to a 'paper panopticon' that oversaw every movement and offence against penal discipline in an open island prison colony (Byard & Maxwell-Stewart, 2018). Combined with civil registrations, censuses, musters and the imperial task of controlling both Indigenous people and the invaders, the British brought clarity and efficiency built on their management of slaves, settlers, colonised people, imperial and commercial military forces, and of course, trade around the globe. Government officials included gifted statisticians who would find themselves given a free hand in the new colonies.

The Australian states and the Commonwealth therefore host a wide array of population data assets that can be mined and linked to build longitudinal and intergenerational datasets. Contemporary administrative data is being automatically linked (Harron et al., 2017), but only the State of Western Australian draws on (albeit limited) historical records. Hand linkage and careful triangulation are required to create reliable historical data. With ingenuity, historical knowledge, good funding and willing volunteers, this has proved possible over the past two decades.

This paper reports on four such projects using both funded research assistants and volunteer genealogists, to build longitudinal datasets built on discrete archives and vital registrations. The projects evolved into a prosopographical practice, where data about individuals identified within a defined universe, are collected into a database that illuminates group characteristics through the creation of an historical collective biography (Charle, 2015). While mostly used for pre-modern and classical historical studies, when utilised in the era of bureaucratic surveillance of populations through vital registration or other records, it is possible to build datasets that are framed by a registered life, while adding multiple other variables for analysis (Pasin & Bradley, 2015).

The four datasets have been derived from collections of records created by institutions: births in a lying-in hospital for the poor (1858–1900), the Aboriginal Protection Board of Victoria, the Tasmanian convict records housed in the Tasmanian Archives, and finally service and pension medical records held in the National Archives of Australia of men who embarked from Victoria for service in World War I. The amplification of variables in cradle-to-grave data through prosopographical practice, has enabled the exploration of key questions in population science about changing life expectancies; race, class and gender penalties; early life exposures; cumulative insults; and wider historical questions about the effect of major historical events and interventions in the private and work lives of people. Small Anglophone former Imperial dominions like Australia and New Zealand offer great opportunities for historical demography, matched in the Anglophone northern hemisphere only by Scotland (Digitising Scotland; Weaver, 2014).

2 THE FOUNDATIONS

The consistent records from all domains in Australia are the civil registrations of births, deaths and marriages, and they commenced in Tasmania in 1838, just a year after Dr. William Farr instituted civil registration in England and Wales (Kippen, 2002). Unfortunately, the penal colony did not replicate Farr's methodology, so that the causes of death are not systematised to an official nosology and very little information about the individuals registered has been retained. It is therefore difficult to link common names without external triangulating information such as from Convict permissions to marry. Those with common names, non-convicts or emancipated convicts cannot be traced.

The gold standard was set by the free colony of Victoria which secured the services of a student of Dr. Farr, William Henry Archer in 1853. Archer was given a free hand and, unlike his mentor in Great Britain, was able to implement the full recommendations of the London Statistical Society for an ideal vital registration scheme, the only place in the world to do so (Hopper, 1986). Victoria's regime remains the most detailed in the Anglophone world, recording for all vital events: birth, marriage, family formation and death; records of multiple generations with ages, in birth order, and whether deceased or living; birth places, and if overseas, time in Australia; and occupations including those of past generations. This is most impressive for birth registrations where mothers had to provide details of all previous births, alive or dead, with names and ages

and in birth order — a level of detail not found in England and Wales until the 1911 census, and even there without the names of the children. Parents' marriages were recorded on birth certificates and corrected later by the registry if deceitful.

However, the sources are not without some challenges. The accuracy of family data for death registrations (see Figure 1) depends on the presence of reliable witnesses in possession of the family history. Births before wedlock are often conveniently forgotten, and dead children not recalled by later witnesses. People who died in institutions often could not provide a biography on admission and many during the 19th century died, as they termed it at the time, 'without friends' — that is with no relatives or acquaintances who knew anything about them. I have found correlations between dying 'without friends' — a fate that convicts dreaded, lest there be no-one to make them a coffin (Karskens, 1998) — and premature death (McCalman, 2009). Therefore, a death certificate provides an accounting of an individual's continuing as well as past relationships and indeed of their quality — was this a family that shared their stories or did not communicate or did not care? Almost 40% of the males who died between 1855 and 1888 in the colony of Victoria died as apparently unattached male immigrants, friendless on foreign soil: a toll of colonisation and migration that is insufficiently recognised.

While the other colonies, including the original settlement of New South Wales, also began civil registration from the 1850s, none replicated the full Victorian regime and neither did New Zealand or Canada. Furthermore, Archer was interested in the population as a phenomenon of immigration and settlement, and still today, the place of birth, time in Australia, and full names and occupations of parents of the deceased, have to be included if known to the witness. The Victorian vital registrations therefore contain a history of migration and intergenerational mobility, even if the most detailed death certificates are completed by winners rather than losers. Most 20th-century migration records have survived and could be linked to family formation and deaths; and 19th-century records of shipping arrivals and departures are preserved. It is therefore relatively easy to link back to British and some Irish records, including censuses, when researching individuals.

Figure 1 Examples of death certificates, Avoca, 1892

SCHEDULE B. DEATHS in the District of Avoca in the Colony of Victoria.				SCHEDULE B. DEATHS in the District of Avoca in the Colony of Victoria.					
No.	When and where Died.	Name and Surname, Rank or Profession.	Sex and Age.	Cause of Death.	Name and Surname of Father and Mother, if known, with Rank or Profession.	Signature of Deputy Registrar, or Date, and Where Registered.	IF BURIAL REGISTERED: When and where buried, or Name of Person who was buried.	When Born, and how long in the Colony, unless otherwise stated.	When, and at what Age, last seen, in order of Birth, their Names and Ages.
1443	January 17 1892	Effie Marshall	Female 14	Meningitis 14 days of illness	Richard Marshall Farmer Sarah Marshall M. H. Lambert	Avoca 1892	Avoca 1892	Avoca 1892	
1444	January 13 1892	Emily Blount	Female 66	Apoplexy 2 months	Henry Blount John Blount M. H. Commonwealth	Avoca 1892	Avoca 1892	Avoca 1892	
1445	January 17 1892	Charles Calnan	Male 48	Cholera 3 days	Charles Calnan Robert Calnan	Avoca 1892	Avoca 1892	Avoca 1892	
1446	February 15 1892	Sarah Brown	Female 68	Apoplexy 15 years	Not known	Avoca 1892	Avoca 1892	Avoca 1892	
1447	March 22 1892	Ann Harvey	Female 32	Apoplexy 2 years	William Harvey Louisa Harvey M. H. Richards	Avoca 1892	Avoca 1892	Avoca 1892	

Note: Victorian Death Certificates for the rural town of Avoca in 1892, giving details of the deceased's parents (when known), birthplace, time in the colonies, marriage and family formation.

3 MELBOURNE LYING-IN HOSPITAL BIRTH COHORT 1857–1900 (LIH BIRTH COHORT 1857–1900)

In the 1990s the vital certificates were indexed and transferred to searchable CDs. In 1998, on completion of a history of Melbourne's Royal Women's Hospital from 1856, based on a remarkable archive of patient midwifery and gynaecological records (McCalman, 1998), the then Professor of Perinatal Medicine, Sean Brennecke, suggested that a study could be made from the midwifery records (see Figure 2) to test the Barker Hypothesis of the foetal origins of adult health. Now it was possible to trace babies born from 1857–1900 to a recorded birth and death, thereby linking their birth weight to life outcomes.

The project, the Melbourne Lying-In Hospital Birth Cohort 1857–1900, traced the life courses of around 8,602 babies for whom full data had survived and whom we could trace to a registered death. We demonstrated that a cradle-to-grave dataset could be built from an existing archive that captured a discrete population. The midwifery records were of high quality as the founding doctors — an Irishman and an Englishman — had both walked the wards in Paris and were determined to apply statistical analysis to their patients and their practice (Warner, 1998). The record book they used had been developed by the great Scottish obstetrician, Sir James Simpson (Simpson's forceps, etc.). The birth record included the mother's age, marital condition, place of birth, parity and length of hard labour, alongside the baby's condition (alive or stillborn), presentation, sex, weight and length. All interventions such as forceps or destructive instruments were noted, the use of chloroform, manual manipulations, as were complications such as obstructed labour, haemorrhage and rupture of the uterus. Maternal deaths were studied carefully. The population in the hospital was overwhelmingly overseas-born, with Australian-born mothers not predominating until the mid 1870s. Birth weights varied according to maternal place of birth, marital status, parity and age, with Scottish and Irish married women having the biggest babies and Victorian and Tasmanian single women, the smallest (McCalman & Morley, 2003).

Few similar records sets have survived for this period world-wide, but the most notable historical studies have been of the Montreal Lying-In Hospital (Ward & Ward, 1984) and in three Norwegian cities, 1860–1984 (Rosenberg, 1988). Our project proved to be the earliest historical investigation of the Barker hypothesis that birth weight might be an indicator of restricted intra-uterine growth and predictor of later adult health, in particular of cardiovascular disease. Working with an English perinatal epidemiologist, Dr. Ruth Morley, we found no relationship between low birth weight and cardiovascular disease, essentially because most small babies, who may have been 'small for dates' from restricted growth, died well before the age of twelve months. This did not invalidate the Barker Hypothesis, but rather revealed limitations of historical birth weight records as an anthropometric measure of early life influences and the need for wider economic and social variables, epigenetic and environmental effects (Almond & Currie, 2011).

Figure 2 Page from the midwifery book, Melbourne Lying-In Hospital, July–August 1886

Note: Pages from the midwifery book for July–August 1886 during an epidemic of Group A Streptococcus. The long pink highlights were maternal deaths; the shorter highlights were of infants traced to a death certificate. The lone yellow highlight is of the only baby who lived to reach adulthood. The birth entries were heavily annotated by a brilliant young medical registrar, Dr. John Dunbar Hooper, who made a detailed study of a hundred consecutive deliveries, nearly all of whom resulted in maternal post-natal infection. He used this evidence to persuade his conservative superiors to adopt antiseptic midwifery, which led to an immediate fall in maternal and neo-natal deaths.

The Melbourne Lying-In Hospital was a charity hospital established for women who did not have a suitable home for their lying in, and around half of the women before the mid-1890s were unmarried, many of them prostitutes. Among the married, desertion and widowhood were common. Hence this was an impoverished population of women who could be divided into varying degrees of being supported or unsupported by a household or a male breadwinner (Morley, McCalman, & Carlin, 2006). Even if there was a father of the child recorded, many of them were, what Jane Humphries has termed 'frail breadwinners': poor providers because of illness, disability, alcoholism or general unemployment (Humphries, 2013).

Birth weight and infant mortality in the LIH Birth Cohort were closely related to the mother's lack of a supporting household of some form. Because we could link the midwifery records to the detailed Victorian birth, marriage and death certificates, as well as later-life records such as military service, criminal offending, divorce and inquests, we had more data for each individual than for those projects in Canada and Norway which had only the hospital birth records (see Table 1 for the basic numbers). We found a steady gradient from those babies whose father was absent from the birth certificate; to those whose father's name was included even though the parents were not yet married; to those whose fathers were married to their mothers but who were unskilled; and finally, skilled fathers. This gradient was in birth weight and in infant mortality, which was extreme among the illegitimate. Likewise, with adult life expectancy, there was a strong class gradient based on geographical location of death, insecure work, family breakdown and criminal activity. Thus, even within a population where all were eligible for admission to a charity hospital, there was a clear gradient of income and life span that was predicated on legitimacy, security of income, housing and stable family life (McCalman, Morley, & Mishra, 2008; McCalman, Morley, Smith, & Anderson, 2011).

Table 1 *Births and death events. Melbourne Lying-In Hospital, 1857–1900*

	Number	Percentage
Registered births	16,290	
Death certificates traces	8,602	52.8
Died < 6 months	4,308	50.1
Died 1–16 years	947	11.0
Died after age 16	3,347	38.9
Lived to age 40	2,958	34.4

4 KOORI HEALTH RESEARCH DATABASE (KHRD)

The potential of this cradle-to-grave life course reconstruction from vital registrations and other records, inspired the leading Indigenous social health academic, Ian Anderson, to suggest a collaboration to reconstitute the Aboriginal population of Victoria. Here the core archival records were those of the Aborigines Protection Board in combination with vital registrations. From these his mother, Sandra Smith, was building genealogies of Victorian Aboriginal people while working at Museum Victoria's Bunjilaka Centre. The number of people discovered by this process was constrained not just by the collapse of the Aboriginal population, but by privacy restrictions on access to vital registrations: deaths can be accessed after thirty years, but marriages only after 70 years and births after a hundred. This left us with the most complete population being from 1870 to 1922, but since we were looking for health transitions, this did capture the critical time period the impact of colonisation on Victorian Aboriginal people. See Table 2 for the final results of the data collection.

The colony of Victoria was the earliest to institute bureaucratic control over Indigenous people in Australia in response to the catastrophic destruction of the Tasmanians and the severity of frontier violence and disease (Boucher & Russell, 2015; Broome, 2005). Working with Dr. Len Smith, who had pioneered the historical demography of Aboriginal Victorians (Smith, 1980), we sought to reconstitute the population from oral genealogy and vital registrations, as the public records rarely noted indigeneity. Aboriginal births, deaths and marriages, were conscientiously recorded by the assistant registrars in each district, even to the extent of naming prominent white settlers who had fathered children with Aboriginal women, but the Aboriginal individuals could only be identified by their family connections (see Figure 3).

Table 2 *Number of reconstituted persons in the KRHD database*

		Males	Females	Unknown
Total KHRD database population	7,900			
Aboriginal	7,405	3,818	3,519	68
Not Aboriginal	495			

Notes by Len Smith:

'Confirmed' individuals:

'Complete' from 1870 to 1922

'Closed' forward and backward:

All known ancestors and siblings of current population

All known descendants of founder population

Most complete date for those born between 1870 and 1922: the periods when the Victorian Aboriginal population was at its lowest, but beginning to recover.

Numbers are small, and analysis undertaken in decadal birth cohorts.

Here again, we were combining archival records with vital registrations of a discrete population of Aboriginal people and their part-descendants. The key questions that Len Smith wanted answered included the extent of initial population collapse under colonisation; how many Aboriginal people remained visible to the state over time; and how many were invisible. Finally, were these 'invisible Aborigines' the source of Aboriginal Victoria's recent recovery? From the time Captain James Cook first mapped the eastern seaboard of the continent in 1770, the population in what is now known as the state of Victoria, crashed from an estimated 60,000 to 15,000 by the time Europeans permanently settled in 1835, to a nadir of 600 in 1900, to no 'full blood' Victorian Aboriginal people alive today since 1993 (Smith et al., 2011). Today, with around 48,000 disclosing Indigenous descent in the 2016 Census, how did the population recover at such a rate when natural increase would have been insufficient?

Various attempts at control had completely failed to protect people until the 1860s when the Board began moving them on to reserves. Around half of those known to the authorities agreed to go on to the reserves, but they tended to be in poor health. There they were encouraged to start farming and to have an education, until pressure from local farmers coveting their land pushed them into greater concentrations, even further away from their respective country. The assumption by the colonial state was that the indigenous population would quietly fade away, however the mid 1880s, a new part-Aboriginal population was beginning to make unwanted demands on the colonial budget. The 1886 'Half-Castes Act' forced those of part descent to leave the reserves and live as 'legal whites': except that the rest of society continued to discriminate against them as 'black'. 'We were too white to be black and too black to be white', they would say (Broome, 2005; McCalman & Smith, 2016).

Through family reconstitution we found a population of 'invisible Aborigines' living outside the surveillance of the Protection Board but remaining connected by kinship to those still living 'under the Board'. Aboriginal women had no entitlement to moral protection against male sexual violence and we found startling evidence of the impact of sexually transmitted infections on fertility: constraining the Aboriginal population from making an early recovery from the impact of colonisation. Likewise, we found similar acquired secondary infertility among convict women (McCalman & Kippen, 2019a). We also found, by comparing the life courses of the poor whites born in the Lying-In Hospital, that the gap between white and black health and survival widened as poor whites made gains with a decline in tuberculosis mortality in particular, while Aboriginal Victorians experienced a rise in tuberculosis and infant mortality that held up until the mid-20th century. Thus, we could demonstrate that the notorious 'gap' that remains a fraught socio-medical and political issue to this day, emerged in response to deliberate government policy of racial management. However, we also found, to our surprise that the known population in Victoria is overwhelmingly descended from those who went on to the reserves in the 1860s and 1870s: that the Board of Protection did save Aboriginal Victoria from complete annihilation. Their concentration in closed reserves at least enabled them to preserve some of their language, their genealogies and their culture (McCalman & Smith, 2016; McCalman, Smith, Anderson, Morley, & Mishra, 2009; McCalman, Smith, Silcot, & Kippen, 2021; Smith et al., 2011).

Figure 3 Death certificates from an Aboriginal reserve in 1928

Note: Death certificate from Lake Tyers Aboriginal Reserve (Victoria) covering 46 days in 1928 when four children died in a population of just over 200 people. The first died at five months from heart failure and gastro-enteritis (probably diagnosed today as failure to thrive); the second was premature and died after 6 hours; the third, aged 13 years, died from a blow on the head inflicted by her sister in a fight; the fourth, a boy aged 1 year, died also from heart failure and toxæmia. The children were all buried by a leader in the community, Jack Mullet. The register was kept on site, hence the ink blots.

5 FOUNDERS & SURVIVORS SHIPS PROJECT (FAS SHIPS COHORT), 1818-1853

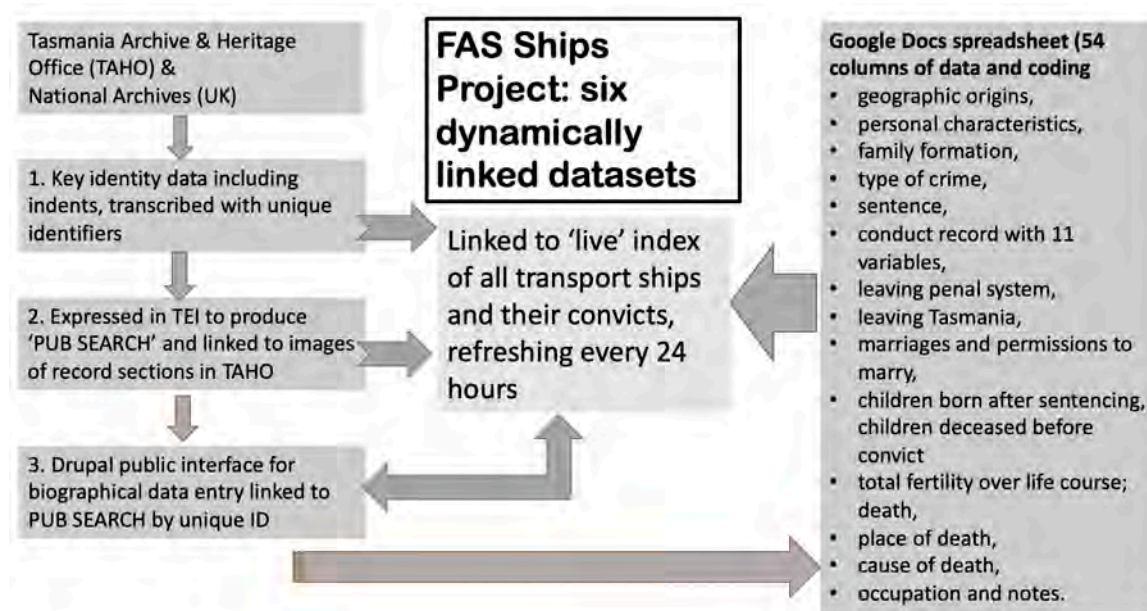
The Melbourne LIH Cohort was enriched by war service records, social welfare case records, criminal records, marriages, electoral rolls and newspaper reports. Likewise, the KHRD dataset was amplified by the addition of every sighting of an individual in the Protection Board archive and newspapers. However, it was with the convict records of Tasmania that it was possible to expand even further the range of research questions and variables.

The seed for 'Founders and Survivors: Australian life courses in historical context' was sown by Rick Steckel of the University of Ohio at a seminar in Columbus in 2005. As an economic historian, his interest in convict records was in any relationship between height, stunting, and vulnerability. The historical demographer Rebecca Kippen now joined the projects and with Hamish Maxwell-Stewart leading from the University of Tasmania, in 2008 a team of demographers, economic historians and social historians, embarked on building a dataset from the about 68,000 individuals who had been transported between 1812 and 1853 and for whom usable records had survived. The technical challenge was the transcribing and structuring of detailed hand-written records of human characteristics and events into a database form that could be used for analysis (see Figure 3). As with LIH and the KHRD, FileMaker Pro supplied the most reliable software for the initial combination of numerical data, coding and text, but to do even that, core records had to be transcribed and allocated to selected categories. This work was undertaken in Tasmania, led by Alison Alexander, and has formed the core data for the complex assemblage of data on those under the Paper Panopticon in combination with the transcriptions donated by Deborah Oxley and David Meredith (Bradley, Kippen, Maxwell-Stewart, McCalman, & Silcot, 2010; McCalman, Smith, Silcot, & Kippen, 2015).

A subsidiary project commenced at the University of Melbourne in 2011 with a systematic reconstruction of convicts' full life courses, including life and family formation, before and after sentence (the FAS Ships Project). Tracing to a recorded death was difficult as so many, particularly Celts, had common names. This project involved a far greater range of data, including the assimilation of full texts of conduct records. The difficult informatics concerned the reconciliation of multiple textual sources for each individual, many of which had small discrepancies with spelling, dates and places. The aim was to use community genealogists as volunteers working on an interactive online platform, and here researchers needed to be able to see all the variants in a given convict's record and to analyse and code their conduct records for later analysis. TEI, or the Text Encoding Initiative (Burnard, 2014), was employed by our system designer, Sandra Silcot, to build a core dataset that amalgamated descriptive data on convicts and then linked that with images from other records in the system: conduct records, description lists, musters, and the indents or embarkation data collected on each convict on arrival, along with physical measurements, and accounts of their crime in their own words and of their families and birth places. Some of these brief flashes of convicts' own speech are very moving, and the descriptions of the families left behind have made it possible to explore the significance of parental or marital loss in offending (Kippen & McCalman, 2016).

Sandra Silcot's data architecture of the FAS ships project was rich and ingenious (see Figure 4). From four separate datasets, a 'biography' could be assembled 'on the fly', so that the life became the sum of the parts of the archive pertaining to that individual. This was held together by a 'live' index which in turn linked to spreadsheets in Google Docs that were completed by the researchers. Further, each line of data for an individual in the spreadsheet had live links to both the core TEI data from the indents and to the biography contributed by the online researchers via a portal run on Drupal. The TEI data then linked to digital images of the relevant original record held in the Tasmanian Archives and Heritage Office: for many of them, part of a page or a whole page in an archived book. The online workplace therefore aggregated individual records from four independent data sites. Users could move back and forth between individuals and groups, original texts and transcripts, with the TEI data recording their lives under the penal system, and the biographies recording their lives before and after sentence as discovered by the researchers. The Google Docs spreadsheet collected the text, enumeration and coding that was later exported into SPSS for analysis (see Figure 5).

Figure 4 Schema of the linked datasets of the FAS Ships project



The project was broken into shiploads of convicts rather than a sample of, for instance, 1:3. The convict voyage had created a mobile community that remained significant in many convicts' personal lives. Each voyage had its own historical context, and most were carefully documented by the ships' surgeons who were required to keep detailed 'sick lists' and were in charge of the convicts both for health and discipline while at sea. Ships surgeons were paid bonuses for landing most of their charges in good condition, and death rates were remarkably low. Moreover, many surgeons were attentive, and most convicts responded with good behaviour. Apart from stormy seas, convict voyages were rather peaceful.

Figure 5 Pages from the three linked datasets: dynamic index, google docs spreadsheet and an individual biography

Agincourt_1844_M226_c33a_b4a372.17_vjs vjs ship progress report

Generated at 2017-08-03T02:10:34:047+10:00 by klanis.

Summary/abbreviated data shown below is generated daily by analyzing a snapshot of the volunteer's spreadsheets and integrating the fat record id with the CCC contribution. Summary details include the CCC death and desc details. Links to the pubsearch data and the linked CCC records are shown. For complete details, follow the links.

Added 2014-04-03: [Quality assurance issues](#).

2) Agincourt_1844_M226_c33a_b4a372.17_vjs

Managed Type	Google Link (sign in)	Convicts Updated	CCC %	% CCC	Editors
data	data	224	28	100	volunte.mcauliffe@gmail.com, jaeerocoulson@179@gmail.com, jaeel6421@gmail.com, nola.beagley@bigpond.com.au, hlorenzani14@gmail.com, 64029168@bigpond.net.au, rtdl@me.com, jeady@researchintransit.com.au, yalves25@gmail.com

Note as at Aug 15 20:51:07 the pubsearch function integrates into the lifecourse (where approved) the ccc contributed data for birth, marriage, death (aif, children not shown yet). These data items are listed in brown.

Legend to symbols:

- (6) CCC entry not yet approved; death & biography is indicated but other details are not available until approved.
- (7) (18) access restricted to staff; volunteer has NOT granted public access
- (90) death NOT found. Death HAS been found for 14065.
- (E) have death, including cause of death
- (S) biographical data recorded
- (B) birth of child
- (D) of descendant
- (72) The death SHOULD be matchable to a Kippen & Ginn death registrations record (to a Tasmanian death in year range 1835-1909). Click the icon to search this dataset using the information about the death which has been provided.

Linked fat records to CCC:

Note that details of CCC entries only present where the entry has been approved by staff.

- Adams, William (Agincourt, 1844) : FAS record pubsearch:23a3830002
- Aldridge, William (Agincourt, 1844) : FAS record pubsearch:23a3830002

Pub Search and CCC have been amalgamated, but were originally linked as distinct data sets

Note: This figure provides screenshots of the main elements of the database: the 'live' index on the left, with its links to the Google Docs spreadsheet (top right) and the data entry web page for volunteers to upload new information on a convict (bottom right). The red arrows are the links between these three datasets.

Breaking the project into 'ships' also made it easier for the volunteers. Many started with the ship or ships bringing their own ancestors; there was always an end in sight, rather than a pitiless dark tunnel of data collection; and many found additional historical material to add to the voyage's story. Their first task, however, was very demanding, even for the experienced family historians. We needed them to be able to read Copperplate handwriting, so they had to be predominantly older people educated before the 1970s. We required them to record and code every piece of relevant data in a long Google Docs spreadsheet: biological and social characteristics: age, height, places of birth and places of conviction, religion, literacy, marital status, family size with occupations and places. Thus, we had multiple variables on the life condition of convict upon entry (see Figure 6).

For the convict's time under sentence researchers had to decipher densely hand-written notes, full of abbreviations, of the new offences and their punishment. Much work had already been done in Tasmania by the Female Convicts Research Centre and the Port Arthur Centre in the art of transcribing convict records. These detailed records can involve three or more hours of concentrated work to decipher and transcribe, which was humanly impossible if we were to reconstitute a large sample of the population. Moreover, once transcribed, the conduct record still needed to be coded for analysis. Therefore, a method of coding directly from the original text was developed to capture a range of offences, reactive behaviours and punishments conceptualised as 'insults' — lashes, days in solitary confinement, days on the treadmill, head shaving, time in chains and hard labour or at the washtubs, time at Port Arthur. The 'crimes' ranged from insolence and refusal to work, to violence, destruction of clothing (related to severe mental illness), drunkenness, sexual offences including same-sex offences, violence and theft. It became clear that drawing conclusions about reactive behaviour had to be tempered by knowledge of agricultural cycles (valuable workers were punished less in harvest periods) and by interpersonal conflicts. Under the 'Assignment' system (1803–1840 for men, 1812–1844 for women), convicts were assigned to private masters or mistresses who were their primary disciplinarians for the duration of their sentence; under the succeeding 'Probation' system they underwent two or more years' controlled labour on probation, remaining under official surveillance until deemed sufficiently reformed to obtain a ticket-of-leave where they could work for wages. Punishment in chains, or in a penal station or the crime classes of the Female Factory, was for secondary offences committed under sentence, and overall, around half of all convicts succeeded in avoiding harsh treatment.

Figure 6 Male conduct record under the Probation system

5503

Transported for *Stealing a Lamb, Gaol Report*
Conduct good, Whingle. Stated this offence.
Stealing a Lamb *Single*

Tried *C. Westmeath 22 July 1850*

Embarcaded *June 1845* Arrived *June 1845*

Roman Catholic. Can Read a little. *Surgeons Report, good*

Trade	Height	Age	Complexion	Head	Hair	Whiskers	Visage	Forehead	Eyebrows	Eyes	Nose	Mouth	Chin	Nat. Place
<i>Sabon</i>	<i>5 ft 10 in</i>	<i>19</i>	<i>Fair</i>	<i>oval</i>	<i>Brown</i>	<i>Nada</i>	<i>Broad</i>	<i>flat</i>	<i>Dark</i>	<i>Grey</i>	<i>Small</i>	<i>Medium</i>	<i>Almond</i>	<i>Co. Wick</i>

Marks *Deeply freckled, very high Cheek, Shank made.*

Period of Labour *2 1/2 Months*

Station of Gang *First by gang*

Class *1st*

Offences and Sentences

Remarks

Emancipated from Gang 1st April 1850

29. 10. 50 52

29. 7. 51 Rec. June 5. 1851/52

142 of 53 Anglin - Long Prisoner - 18 months hard labour

1st of 14 of 1853 19 April 55 Johnston 8 1/2 months

in prison to work. Six months hard labour 1815 7/1855

Act of Freedom 28 January 1836

This difficult work required extensive training with workshops, backed by online and paper manuals, but the volunteers rose to the challenge and worked with great enthusiasm and accuracy. The university research team included paid research assistants who checked new entries before they were accepted and provided daily support to the wider team. The volunteers' morale was maintained with regular week-end lunches and an illustrated online magazine was produced three times a year for four years, with contributions from the volunteers and the academic staff (see Figure 7).

Around sixty volunteers, many from other states and one from overseas, worked on the project over a period of four years and produced nearly 25,000 biographies out of 68,000 convicts (Kippen & McCalman, 2016). Our work was made far easier by a close collaboration with the Female Convict Research Centre in Hobart, which also uses volunteer labour (FCRC).

The Female Convict Research Centre has been an essential partner in all this work. They pioneered using volunteers for the systematic transcription of the female convict records. However, they were short of resources. By forming a partnership, we were able to resource them with access to paid death certificates and staff time, while we shared records and training. They are now very close to completing the records of all women convicts sent to Tasmania and this is available to all who register with the Centre online.

Given that most former convicts were anxious to become invisible upon release, and many had common names, our volunteers were remarkably successful in tracing enough men and women, before and after sentence (see Table 3).

Figure 7 No 10., April 2012 of 'Chainletter'

FOUNDERS & SURVIVORS



Chainletter

News from the Founders & Survivors project at the Universities of Tasmania, Melbourne, Flinders, Monash and the Australian National University

A Tri-annual Newsletter
Issue No. 10 April 2012

<p>Ships Projects</p> <p>We report on the wonderful progress of the Ships Projects and the plan for the next twelve months. In particular we now need to start work on the Women's Ships. Page 2</p>	<p>Next Workshops</p> <p>Our next workshops will be in Hobart & Launceston on 2 & 3 June to launch the women's ships projects. On 23 June, we plan a day-long workshop at Melbourne University. Page 2</p>	<p>Ships Projects Stories</p> <p>Steve Rhodes returns with two life stories from the Southworth 1834. Pages 3 to 5</p> <p>And Glad Wishart reports on new connections. Page 12</p>	<p>New Book</p> 	<p>Surgeon Superintendent</p> <p>James Bradley discusses new research on the medicine of convict voyages and the role of the surgeon superintendent. Page 7</p>	<p>Research Reports</p> <p>Claudine Chionh reports on the recent Digital Humanities conference and Janet McCalman & Rebecca Kippen on their paper at the European Social Science History Conference in Glasgow. Page 11</p>
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Editorial

We are now able to establish some deadlines for our project, as our funding will finish at the end of 2013.

The Australian Research Council has been very generous since 2007, but we know that we cannot expect further funding after the current grants are exhausted.

Sadly this means that the Founders & Survivors website will need to be archived, probably by the National Library of Australia's Pandora. This will keep it open for people to consult, but it will no longer be interactive. This means that new entries will no longer be permitted and communications maintained. December 2013 will mark the end of *Chainletter* as well.

Chainletter No. 10 April 2012

However, it is not all sad news. We will find ways to publish results of the project online, the collection of convict biographies will be available, and we hope that the detailed research database will be accessible to researchers through Monash and Melbourne universities.

We have just received another grant from the Australian National Data Service (ANDS) to enable the data to remain useful to researchers in perpetuity.

We cannot, however, obtain funding that will employ the staff to run the service for the public. Funds for libraries, museums and the arts are tight and are likely to become even tighter in the near future.

In the light of this, we are thinking about online publishing as a means of continuing the research. And we have been approached to provide material that can be used for schools nationwide. Roar Films in Tasmania have been developing film and media using the project and we have contributed to a feature-length movie on Ikey Solomon.

Our research strategy therefore is to try and finish most of the ships by July 2013, so we have time to clean up the data and tidy loose ends.

In Volunteers' Corner we outline the plans for the next stage: women's ships and key men's ships that we need researched to complete the project.

Our prosopographical data collection has enabled us to undertake multivariate analysis with wide range of variables and many of our starting hypotheses, which looked eminently reasonable historically, were not sustained by the analysis: height and literacy, for instance, were not significant in convicts' life outcomes (stunting was expected to reduce lifespan; literacy to extend it). Similarly, the harsh physical punishments inflicted under sentence did not shorten lives whereas days spent in solitary confinement did (Kippen & McCalman, 2015). Women were far less resilient than men, entering the penal system with major psychological and physical vulnerabilities that disposed them to alcoholism, difficulties with trust and intimacy, and a huge burden of acquired infertility from sexually transmitted infections. Convict men who survived sentence in fact lived longer than their social peers in Great Britain and Ireland; convict women did worse than their peers back home. As we traced their lives before and after sentence, the factors that emerged were new to convict historiography: the important push factor was the fracturing of families and households, leaving children and young people without emotional and material support (Kippen & McCalman, 2018). And the most significant determinant of lifespan was the crime economy of the convicts' birthplaces: seaports, being especially dangerous for women — their mothers — were the worst places to be born; rural villages even in Ireland, the best. This resonates with modern research on toxic stress in utero and early life, foetal alcohol syndrome and neglect, damaging children's cognitive and social early development, scarring them for life (McCalman & Kippen, 2019b). Likewise, the higher life expectancy of male emancipists compared to the peers left behind them, suggested that the better diet and climate in Australia conferred biological mid-life benefits, as understood in life course epidemiology (Kuh, Ben-Shlomo, Lynch, Hallqvist, & Power, 2003).

Table 3 Percentages of the sampled persons that could be traced to death. Founders & Survivors Ships Project, 1818–1853

Sex	%	Year of arrival	%
Male	44.3	1812–1829	50.6
Female	49.1	1830–1839	46.8
		1840–1842	44.8
Country of birth	%	1843–1845	42.8
England	46.7	1846–1849	46.5
Ireland	43.6	1850–1853	45.0
Scotland	45.1		
Other British	46.0	Age at arrival (years)	%
Other	44.3	7–19	39.0
Not recorded	61.4	20–24	43.6
		25–29	46.9
Place of birth	%	30–39	48.8
Village	46.2	40+	56.3
Town	47.4	Not recorded	57.8
Industrial urban	44.2		
Port cities	41.2	Offences under sentence	%
London	40.9	(exclude convicts who died	
Other country	43.8	within five years of arrival	
Not recorded	61.7	None	47.8
		1–2 offences	43.7
		3–5 offences	41.6
		6+ offences	38.6
		Constant	41.0
		Not recorded	21.3

Explanation: The Ships Project included 124 ships with 16,953 males and 7,783 females. This is a clustered sample from an estimated number of 68,000 persons that are included in the database of the Founders & Survivors project.

6 DIGGERS TO VETERANS: RISK, RESILIENCE AND RECOVERY IN THE FIRST AUSTRALIAN IMPERIAL FORCE (AIF) IN WORLD WAR I

This final project pulls the cradle-to-grave studies into the 20th century, taking a systematic sample of men who embarked for overseas service in World War I in the State of Victoria as a discrete population for a demographic prosopography. The purpose was to examine the life course effects of war service exposures, and the early life factors that influenced the impact of the insults of war service on individual lives. Australia has retained a wide range of records that makes such a study feasible: military service records have all survived, whereas those for World War I in the United Kingdom and the United States have not. These are digitised and openly available from the National Archives of Australia. The medical examinations for disability pensions have also been preserved and can be accessed on individual request from the same archives. Linkage to civilian birth, marriage and death records is easy in Victoria, and whereas New Zealand and Canada have many comparable records available, in neither case is it currently possible to link to a registered death in

the public domain (Wilson, Clement, Bannister, & Harper, 2014). In Victoria, deaths have only a thirty-year embargo so that virtually all the veterans who died in Victoria could be traced to a detailed death certificate.

We may lack census returns, but once you are reconstructing lives in the 20th century, there are Commonwealth electoral rolls which at least reveal cohabiting adults, nearby relatives and occupations, even if they are often too general to be particularly useful: e.g. 'public servant' could cover anyone from a train driver to a senior administrator. However, the electoral roll gives residential address, which remains the best indicator of socio-economic status in combination with occupation — and since they were created for elections, they were more frequent than the censuses. Those with unstable living arrangements can be easily discerned, whereas automated linkage at least of the US censuses appears to privilege the settled over the unsettled. Voting has been compulsory in Australia since 1922 and from the start of the Federation in 1901, the electoral office has always enrolled people without a fixed address — initially, at the insistence of the Labor Party, because so many rural workers were itinerant yet unionised. Female suffrage also began in the Commonwealth in 1902. These rolls, along with a range of other government records like Government Gazettes and some municipal rate books, have been digitised by Ancestry, while State Government records of criminals, neglected and delinquent children being made Wards of the State, divorces, rural land titles, are online. War Service rural settlers, inquests, probates and wills are also readily available online at no cost. Finally, the National Library of Australia pioneered the digitising of all historical Australian newspapers in its collection: from major metropolitan newspapers, to the provincial and niche — such as German-language newspapers from the Barossa Valley in South Australia. These are also freely searchable online as part of a national 'TROVE' of all known publications and images relating to people, places or events. British and American digitised newspapers are behind paywalls.

These outstanding online resources have revolutionised Australian historical research and vastly expanded the capacity to construct rich genealogies and prosopographies. The Diggers to Veterans project used two teams of volunteers:

- (1) an historical team of genealogists who
 - traced the servicemen's families, often back two or three generations with family sizes, frequency of infant and child mortality;
 - used newspaper, inquests, criminal and state ward records to record evidence of family dysfunction, father's occupations, premature deaths of parents; mental illness or suicide — all of which proved to be risk factors for those in military service;
 - analysed and coded the soldier's service record for wounds, sickness, conduct offences, valour awards, shell shock, gassing, and actual exposure to combat;
 - reconstructed the veterans' life after the war using electoral rolls, marriage and birth records and notices, newspaper records and the death certificate (see Figure 7).
- (2) a medical team of retired health personnel: doctors, a former professor of physiotherapy and a psychiatric nurse. All had had clinical experience with veterans of both World Wars.
 - This team worked through the medical and pension files of those veterans who engaged with the Department of Repatriation (now Veterans Affairs).
 - They assessed the seriousness of the claims, and quality of the diagnosis, the state of the man's health over time, and the treatment. Much of the real story of these men's health after military service was concealed by the fact that no-one asked them about smoking until the 1970s. Alcoholism was common; suicide around the same as the wider community. Largely, however, their health declined in concert with ageing and the wider population.

This project is not finished, and we have published only a preliminary demographic analysis of the first half of the sample, who are the persons with the longest overseas service as the sample proceeds by time of enlistment (see Table 4). The most important determinants of life span were personal characteristics such as early life conditions and socio-economic status, and not war exposures apart from a small number of severe insults (McCalman, Kippen, McMeeken, Hopper, & Reade, 2019). The Union Army data led by Dora L. Costa (Costa, 2012; Costa, DeSomer, Hanss, Roudiez, Wilson, & Yetter, 2017; Costa, Kahn, Roudiez, & Wilson, 2018) over many years has been an inspiration to Diggers to Veterans, and it is hope that the Australian data will be enhanced and utilised by many researchers over the coming years.

Figure 8 Example of service record of deployments and medical admissions

ANDERSON		Ernest George.	6456	21st Bde 5th
Surname		Other Names.	Regimental No.	Unit.
WAR GRATUITY SCHEDULE		PURPORT.	3rd M.D	AUTHORITY.
Embarked at Melbourne Victoria on H.M.A.T. "A 4 BUBIPINS" on 11/9/15		Pte <i>Nesta</i>		
5.8.17 Adm. to Tooting Mil/ Hosp. Epididymitis		IB 288.3.17		
13.3.17. Proceeded overseas to France from Folkestone. ex 2nd. Tng. Btn. Durrington/		Lon. 19/5-17		
28.0.17. Trans from Tooting M Hosp. to 3rd Aux Hosp. Epididymitis.		IB 307/3-17		
30-13.9.17. Dis. from 3rd Aux Hosp. Fur & rep. P Downs. Epididymitis		IB 307/4-17		
5.8.17. Adm. Tooting Mil. Hosp. EPIDIDYMITIS.		Lon. 61/2-17		
13.7.17. Rejoined unit from Hosp.		BEFO. 43/3-17		
7.7.17. To Hosp. SICK		BEFO. 43/4-17		
3.8.17. Emb. on H.S. "Grantully Castle" for England. EPIDIDYMITIS.		BEFO. 49/2-17		
P.T.O..				

National Archives of Australia

NAA: B2455, ANDERSON E G

PURPORT.	AUTHORITY.
4/8/17 Embkd per H.S. "Grantully Castle" from France for return to England. Epididymitis)	LON. 68/1-17
13/9/17 Crime. Perham Downs. A.W.L. 13/9/17 -14/9/17. Award 3 days C.C.. Forfts 2 days Pay. P.B.No. 126008/16	LON. 76/3-17
11.8.18 Adm to L. of C. Hpl from wounds.	IL BEFO. 44/4-18
9.8.18 Wounded in action.	IL BEFO. 44/14-18
10.11.17 Proc O/seas to France via S'ton ex O/seas Tng Bn.	LON 87/3-17
17.7.17. To Hosp., sick.	BEFO. 78/4-17
17.11.17. Rejoined unit from Hosp. (England)	BEFO. 80/4-17
20.3.17. T.O.S. of 5th. Bn. AIF. from 21/5th. Bn. AIF. (LB)	BEFO. 31/7-17
Illegally absent from 1.10.18. (98-5/19) LK	BEFO 55/1918
30/9/18 Diso to P/B&E* from L. of C. Hosp	(MC) BEFO 50/14-18
11.8.18 Adm 12th USA Gen. Hpl Wounded (Sts 5th Btn.) (MH)	IB 548/4-18

National Archives of Australia

NAA: B2455, ANDERSON E G

Note: Service record of an infantryman, pianist in civilian life with sexually transmitted medical problems that kept him out of line for most of his service. Wounded in right hand August 1918, which healed. Four times absent without leave, totalling 66 days. In civilian life he became an alcoholic, a heavy smoker, a convicted bigamist and could not keep a stable residential address. Worked intermittently as a barman. Died aged 67 of acute general peritonitis. Last residence was church refuge for homeless men.

Table 4 *Numerical overview of the Diggers to Veterans sample*

Estimated number who embarked from Victoria	85,000	
Digger to Veterans sample total	11,980	
1:4 who embarked October 1914 – March 1915	2,756	
1:8 who embarked April 1915 – 7 November 1918	9,276	
Sample so far analysed and published	6,183	%
Died 1915–1918	1,387	22.4
Deaths not traced for men who survived war service	607	9.8
Post-war mortality traced for	4,189	67.8

6 CONCLUSION

The disciplinary driver of all these projects has been historical rather than demographic and the questions asked of the data have arisen from the wider Australian historiography. The prosopographical method has yielded valuable data that have over-turned long-held assumptions about the extent and impact of childhood deprivation in Australia and in the home countries of transported convicts. It has answered some key population questions about the impact of colonisation on Indigenous people, in particular on their fertility. These datasets, however, are only a start and I hope that other scholars and population scientists will expand and mine them for many years to come, asking more fine-grained questions about the life courses of Australian people.

This integration of historical and demographic expertise has borne fruit in a more nuanced analysis of social class based on historical geography and economic history. The role of insecure work has not been sufficiently acknowledged in the past, especially in Australia where perceptions of poverty have been clouded by ideas of 'the working man's paradise' and high living standards by the end of the 19th century, severe depressions notwithstanding. In convict studies, analysis of birth places was crudely divided between urban and rural, while not allowing for the differences in female employment between port cities, industrialised cities and service economies, regional market towns and villages. These differences were crucial, we found, in the early life conditions of convicts whose mothers were subject to the violence, alcohol and sexual abuse of port cities. Here historical criminology pointed to the significant differences in these early environments.

The other driver in these four projects has been the evolving thinking in historical life-course epidemiology and demography. This work began with James C. Riley who worked on the frailty and insult accumulation hypotheses using the measurement of sickness episodes, arguing in 2003, for instance, that the cumulative health burdens of reproduction compromised women's life course outcomes (Riley, 2003). Earlier his work on cumulative insults stimulated Diana Kuh and George Davey Smith in their early formulation of life course epidemiology (Kuh & Davey Smith, 1993) where they have added chains of risk extending from early life influences to follow people through time, using the big British birth cohorts (Kuh et al., 2003; Wadsworth & Kuh, 1997). While not able to engage with this vast literature, it was none the less intriguing to conceptualise and compare the convicts and the World War I soldiers as cohorts exposed to relatively short-lived stress regimes within closed societies or total institutions: penal servitude and military service. The experience of both cohorts was closely recorded, some of the behaviours were similar such as bolting, violence, insubordination and drunkenness, and both regimes were exacted on younger people who could be expected to display faster recovery from physical trauma. The results, so far, are interesting where psychological insults (solitary confinement, battle stress or neurasthenia and shell shock) were more damaging in the long term than physical insults (flogging, wounds), but in both populations early life experiences remained important in resilience, as did later life social position. Male convicts who married wisely, drank little, found a way to make a living and rear children, did better than if they had not been transported; middle-class veterans who may have had severe wounds and shell shock still did better in life than those whose parents were unstable and poverty-stricken. In both populations, personal characteristics proved more significant than their respective exposures to stress and insult. This is only the beginning of this work, and the Diggers to Veterans dataset should yield fine-grained data in future years.

These four projects illustrate the long-term potential of data linkage in Australia based on vital registration. Many of us hope that a change in government attitudes to university funding in general and social sciences research in particular, may one day enable the digitisation and linkage of the states' and territories' vital registrations, and their further linkage to Medicare data (the only nation-wide personal data collection apart from taxation records). They could then be linked to other welfare, economic and historical data. Only the state of Western Australia has embarked on such a task, with its Data Linkage WA (<https://www.datalinkage-wa.org.au/>) which provides the best socio-medical data in Australia, especially of Indigenous citizens. However, its vital registrations linkage only goes back to 1944. The ambition is to begin with the first vital registrations and build a dataset of a colonising society and its Indigenous people from the 1830s to the present day. The COVID-19 pandemic and its economic impacts, especially on university research incomes, have stalled these plans. The future of historical population data depends now on government support to enable us to surmount the hurdle of transcription and linkage. If we can build a national historical register of the people on this continent, linked to multiple other data from all domains of administration, health, economics and society, we could produce a dataset of international value that is scalable and relevant to a world reshaped by international migration over the past two hundred years.

ACKNOWLEDGEMENTS

I wish to thank my statistician colleagues Len Smith, Rebecca Kippen, Ruth Morley and Gita Mishra for their advice and contributions. It was their work, along with our research colleagues and volunteers, who made all this possible.

The research for this paper was funded by the Australian Research Council and the Australian National Data Service.

REFERENCES

- Almond, D. & Currie, J. (2011). Killing me softly: The fetal origins hypothesis. *Journal of Economic Perspectives*, 25(3), 153–172. doi: [10.1257/jep.25.3.153](https://doi.org/10.1257/jep.25.3.153)
- Boucher, L., & Russell, L., (Eds.) (2015). *Settler colonial governance in nineteenth century Victoria*. Acton, Canberra: ANU Press and Aboriginal History.
- Bradley, J., Kippen, R., Maxwell-Stewart, H., McCalman, J., & Silcot, S. (2010). Research note: The founders and survivors project. *The History of the Family*, 15(4), 467–477. doi: [10.1016/j.hisfam.2010.08.002](https://doi.org/10.1016/j.hisfam.2010.08.002)
- Broome, R., (2005). *Aboriginal Victorians: A history since 1800*. Crows Nest, Sydney: Allen & Unwin.
- Burnard, L. (2014). *What is the Text Encoding Initiative? How to add intelligent markup to digital resources* (No. 3). Marseille: OpenEdition Press. doi: [10.4000/books.oep.426](https://doi.org/10.4000/books.oep.426)
- Byard, R. W., & Maxwell-Stewart, H. (2018). The potential forensic significance of convict archives from Van Diemen's Land, 1820–1877. *Forensic Science, Medicine and Pathology*, 14, 127–132. doi: [10.1007/s12024-017-9913-2](https://doi.org/10.1007/s12024-017-9913-2)
- Charle, C. (2015). Prosopography (collective biography). *International encyclopedia of the social and behavioral sciences* (2nd edition) (pp. 256–260). doi: [10.1016/B978-0-08-097086-8.62146-3](https://doi.org/10.1016/B978-0-08-097086-8.62146-3)
- Costa, D. L. (1993). Height, weight, wartime stress, and older age mortality: Evidence from the Union Army records. *Explorations in Economic History*, 30(4), 424–449. doi: [10.1006/exeh.1993.1018](https://doi.org/10.1006/exeh.1993.1018)
- Costa, D. L. (2012). Scarring and mortality selection among Civil War POWs: A long-term mortality, morbidity and socioeconomic follow-up. *Demography*, 49(4), 1185–1206. doi: [10.1007/s13524-012-0125-9](https://doi.org/10.1007/s13524-012-0125-9)
- Costa, D. L., DeSomer, H., Hanss, E., Roudiez, C., Wilson, S. E. & Yetter, N. (2017). Union Army veterans, all grown up. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 50(2), 79–95. doi: [10.1080/01615440.2016.1250022](https://doi.org/10.1080/01615440.2016.1250022)
- Costa, D. L., Kahn, M. E., Roudiez, C., & Wilson, S. (2018). Data set from the Union Army samples to study locational choice and social networks. *Data in Brief*, 17, 226–233. doi: [10.1016/j.dib.2017.12.007](https://doi.org/10.1016/j.dib.2017.12.007)

- Digitising Scotland. (n.d.). Retrieved from <https://digitisingscotland.ac.uk/>
- FCRC: Female Convict Research Centre, Tasmania. (n.d.). Retrieved from <https://www.femaleconvicts.org.au>
- Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimae, M., Barreto, M. L., Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data & Society*, 4(2), 1–12. doi: [10.1177/2053951717745678](https://doi.org/10.1177/2053951717745678)
- Hopper, J. (1986). The contribution of W. H. Archer to vital statistics in the colony of Victoria. *Australian Journal of Statistics*, 28(1), 124–137. doi: [10.1111/j.1467-842X.1986.tb00590.x](https://doi.org/10.1111/j.1467-842X.1986.tb00590.x)
- Hull, T. H. (2007). The strange history and problematic future of the Australian census. *Journal of Population Research*, 24, 1–22. doi: [10.1007/BF03031876](https://doi.org/10.1007/BF03031876)
- Humphries, J. (2013). Childhood and child labour in the British industrial revolution. *The Economic History Review*, 66(2), 395–418. doi: [10.1111/j.1468-0289.2012.00651.x](https://doi.org/10.1111/j.1468-0289.2012.00651.x)
- Karskens, G. (1998). Death was in his face: Dying, burial and remembrance in early Sydney. *Labour History*, 74, 21–39. doi: [10.2307/27516551](https://doi.org/10.2307/27516551)
- Kippen, R. (2002). An indispensable duty of Government: Civil registration in nineteenth-century Tasmania. *Tasmanian Historical Studies*, 8(1), 42–58. Retrieved from <http://hdl.handle.net/1885/43120>
- Kippen, R., & McCalman, J. (2015). Mortality under and after sentence of male convicts transported to Van Diemen's Land (Tasmania), 1840–1852. *The History of the Family*, 20(3), 345–365. doi: [10.1080/1081602X.2015.1022198](https://doi.org/10.1080/1081602X.2015.1022198)
- Kippen, R. & McCalman, J. (2016). Crowdsourcing convict life courses, or the value of volunteers in the age of digital data. In K. Matthijs, S. Hin, J. Kok & H. Matsuo (Eds.), *The future of historical demography. Upside down and inside out* (pp. 201–204). Leuven/Den Haag: Acco.
- Kippen, R., & McCalman, J. (2018). Parental loss in young convicts transported to Van Diemen's Land (Tasmania), 1841–53. *The History of the Family*, 23(4), 656–678. doi: [10.1080/1081602X.2018.1513855](https://doi.org/10.1080/1081602X.2018.1513855)
- Kuh, D., & Davey Smith, G. (1993). When is mortality risk determined? Historical insights into a current debate. *Social History of Medicine*, 6(1), 101–123. doi: [10.1093/sochis/6.1.101](https://doi.org/10.1093/sochis/6.1.101)
- Kuh, D, Ben-Shlomo, Y., Lynch, L., Hallqvist, J., & Power, C. (2003). Life course epidemiology. *Journal of Epidemiology & Community Health*, 57(10), 778–783. doi: [10.1136/jech.57.11.914-a](https://doi.org/10.1136/jech.57.11.914-a)
- McCalman, J. (1998). *Sex and suffering: Women's health and a women's hospital: The Royal Women's Hospital, Melbourne 1856–1996*. Carlton: Melbourne University Press.
- McCalman, J. (2009). To die without friends: Solitaries, drifters and failures in a New World society. In G. Davison, P. Jalland & W. Prest (Eds.), *Body and mind: Historical essays in honour of F. B. Smith* (pp. 173–194). Melbourne: Melbourne University Press.
- McCalman, J., & Kippen, R. (2019a). "A wise provision of nature for the prevention of too many children": Evidence from the Australian colonies. In S. Szreter (Ed.), *The hidden affliction: Sexually-transmitted infections and infertility in history* (pp. 279–302). New York: University of Rochester Press.
- McCalman, J. & Kippen, R. (2019b). The life-course demography of convict transportation to Van Diemen's Land. *The History of the Family*, 25(3), 432–454. doi: [10.1080/1081602X.2019.1691621](https://doi.org/10.1080/1081602X.2019.1691621)
- McCalman, J., Kippen, R., McMeeken, J., Hopper, J., & Reade, M. (2019). Early results from the 'Diggers to Veterans' longitudinal study of the Australian men who served in the First World War: Short- and long-term mortality of early enlistees. *Historical Life Course Studies*, 8, 52–72. doi: [10.51964/hlcs9307](https://doi.org/10.51964/hlcs9307)
- McCalman, J., & Morley, R. (2003). Mother's health and babies' weights: The biology of poverty at the Melbourne Lying-In Hospital, 1857–83. *Social History of Medicine*, 16(1), 39–56. doi: [10.1093/shm/16.1.39](https://doi.org/10.1093/shm/16.1.39)
- McCalman, J., Morley, R., & Mishra, G. (2008). A health transition: Birth weights, households and survival in an Australian working-class population sample born 1857–1900. *Social Science & Medicine*, 66(5), 1070–1083. doi: [10.1016/j.socscimed.2007.11.040](https://doi.org/10.1016/j.socscimed.2007.11.040)
- McCalman, J., Morley, R., Smith, L., & Anderson, I. (2011). Colonial health transitions: Aboriginal and 'poor white' infant mortality compared, Victoria 1850–1910. *The History of the Family*, 16(1), 62–77. doi: [10.1016/j.hisfam.2010.09.005](https://doi.org/10.1016/j.hisfam.2010.09.005)
- McCalman, J. & Smith, L. (2016). Family and country: Accounting for fractured connections under colonization in Victoria, Australia. *Journal of Population Research*, 33(1), 51–65. doi: [10.1007/s12546-016-9160-5](https://doi.org/10.1007/s12546-016-9160-5)

- McCalman, J., Smith L., Anderson, I., Morley, R., & Mishra, G. (2009). Colonialism and the health transition: Aboriginal Australians and poor whites compared, Victoria, 1850–1985. *The History of the Family*, 14(3), 253–265. doi: [10.1016/j.hisfam.2009.04.005](https://doi.org/10.1016/j.hisfam.2009.04.005)
- McCalman, J., Smith, L., Silcot, S. & Kippen, R. (2015). Building a life course dataset from Australian convict records. Founders and survivors: Australian life courses in historical context, 1803–1920, In G. Bloothoof, P. Christen, K. Mandemakers & M. Schraagen (Eds.), *Population Reconstruction* (pp. 285–298). Heidelberg: Springer.
- McCalman, J., Kippen, R., Smith, L., & Silcot, S. (2021). Origins of 'the gap': Perspectives on the historical demography of Aboriginal Victorians. *Journal of Population Research*, 38(1), 53–69. doi: [10.1007/s12546-020-09253-x](https://doi.org/10.1007/s12546-020-09253-x)
- Morley, R., McCalman, J., & Carlin, J. (2003). Trends in birthweight between 1857 and 1883, in Melbourne, Australia. *Paediatric and Perinatal Epidemiology*, 17(3), 236–243. doi: [10.1046/j.1365-3016.2003.00500.x](https://doi.org/10.1046/j.1365-3016.2003.00500.x)
- Morley, R., McCalman, J., & Carlin, J. B. (2006). Birth weight and coronary heart disease in a cohort born 1857–1900 in Melbourne, Australia. *International Journal of Epidemiology*, 35(4), 880–855. doi: [10.1093/ije/dyl032](https://doi.org/10.1093/ije/dyl032)
- Pasin, M, & Bradley, J. (2015). Factoid-based prosopography and computer ontologies: Towards and integrated approach. *Digital Scholarship in the Humanities*, 30(1), 86–97. doi: [10.1093/llc/fqt037](https://doi.org/10.1093/llc/fqt037)
- Riley, J. (2003). Did mothers begin with an advantage? A study of childbirth and maternal health in England and Wales, 1778–1929. *Population Studies*, 57(1), 5–20. doi: [10.1080/0032472032000061695](https://doi.org/10.1080/0032472032000061695)
- Rosenberg, M. (1988). Birth weights in three Norwegian cities, 1860–1984: Secular trends and influencing factors. *Annals of Human Biology*, 15(4), 275–288. doi: [10.1080/03014468800009751](https://doi.org/10.1080/03014468800009751)
- Smith, L. R. (1980). *The Aboriginal Population of Australia*. Acton, Canberra: ANU Press.
- Smith, L., McCalman, J., Anderson, I., Smith, S., Evans, J., McCarthy, G., & Beer, J. (2011). Fractional identities: The political arithmetic of Aboriginal Australians. In P. Axelsson & P. Sköld (Eds.), *Indigenous peoples and demography: The complex relationship between people and statistics* (pp. 15–32). New York & Oxford: Berghahn Books.
- Wadsworth, M. E. J., & Kuh, D. J. L. (1997). Childhood influences on adult health: A review of recent work from the British 1946 national birth cohort study, the MRC National Survey of Health and Development. *Paediatric and Perinatal Epidemiology*, 11(1), 2–20. doi: [10.1046/j.1365-3016.1997.d01-7.x](https://doi.org/10.1046/j.1365-3016.1997.d01-7.x)
- Ward, W. P., & Ward, P. C. (1984). Infant birth weight and nutrition in industrializing Montreal. *The American Historical Review*, 89(2), 324–345. doi: [10.2307/1862555](https://doi.org/10.2307/1862555)
- Warner, J. H. (1998). *Against the spirit of system: The French impulse in nineteenth-century American medicine*. Princeton, New York: Princeton University Press.
- Weaver, J. C. (2014). *Sorrows of a century: Interpreting suicide in New Zealand, 1900–2000*. Montreal: McGill-Queen's University Press.
- Wilson, N., Clement, C., Summers, J. A., Bannister, J., & Harper, G. (2014). Mortality of first World War military personnel: Comparison of two military cohorts. *British Medical Journal*, 349(g7168). doi: [10.1136/bmj.g7168](https://doi.org/10.1136/bmj.g7168)

ARCHIVED DATASETS

Inquiries about access to the LIH, Ships Cohort, Convicts to Diggers and Diggers to Veterans datasets should be directed to Associate Professor Rebecca Kippen, Monash School of Rural Health, Bendigo, Victoria, 3550, Australia, rebecca.kippen@monash.edu.

Access to the KHRD is controlled by the Indigenous Data Network, University of Melbourne and permission should be sought from Professor Marcia Langton, m.langton@unimelb.edu.au.

The Diggers to Veterans and KHRD data must be de-identified under the original conditions of access.

Kippen, R., Maxwell-Stewart, H., Alahakoon, D., Bradley, J., Dharmage, S., Inwood, K., ... McCalman, J. (2017). *Convicts and diggers: A demography of life courses, families and generations* [Dataset on war service and ancestry of 1,873 Tasmanian men who enlisted for WWI]. doi: [10.4225/03/59ed3568a4305](https://doi.org/10.4225/03/59ed3568a4305)

McCalman, J., & Kippen, R. (2017). *Founders and survivors: Life course ships project* [Data set of 25,000 convicts transported to Tasmania]. doi: [10.4225/03/59ed402437518](https://doi.org/10.4225/03/59ed402437518)

McCalman, J., & Kippen, R. (2018). Diggers to veterans database [Database]. Retrieved from https://figshare.com/articles/Diggers_to_Veterans_database/5936899

McCalman, J., Morley, R. (2018). *Lying-In Hospital (LIH) cohort, 1857–1900* [Data set]. Retrieved from <https://figshare.com/s/69d0087383df3844bf10>

McCalman, J., Smith, L., Silcot, S., & Kippen, R. (2018). *Koori Health Research Database (KHRD)* [Database]. Retrieved from https://figshare.com/articles/Koori_Health_Research_Database/5936884

Reconstructing a Longitudinal Dataset for Tasmania

Trudy Cowley	Monash University
Lucy Frost	University of Tasmania
Kris Inwood	University of Guelph
Rebecca Kippen	Monash University
Hamish Maxwell-Stewart	University of New England
Monika Schwarz	Monash University
John Shepherd	University of New England
Richard Tuffin	University of New England
Mark Williams	University of Tasmania
John Wilson	University of South Australia
Paul Wilson	University of Tasmania

ABSTRACT

This article describes the formation of The Tasmanian Historical Dataset a longitudinal data resource spanning the 19th and early 20th century. This resource contains over 1.6 million records drawn from digitised prison and hospital admission registers, military enlistment papers, births, deaths and marriages, census and muster records, arrival and departure lists, bank accounts and property valuations, maps and plans and meteorological observations. As well as providing an account of the many different sources that have been digitised coded and linked as part of this initiative, the article outlines current and past research uses to which this data has been put. Further information on tables and key variables is provided in an appendix.

Keywords: Longitudinal historical datasets, History of crime, History of health, Life course history, Historical demography, Historical GIS

DOI article: <https://doi.org/10.51964/hlcs10912>

© 2021, Cowley, Frost, Inwood, Kippen, Maxwell-Stewart, Schwarz, Shepherd, Tuffin, Williams, Wilson, Wilson
This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

The Tasmanian Historical Dataset is a longitudinal data resource spanning the 19th and early 20th century. It includes information on all Tasmanian recorded births and marriages in the period 1803–1899 and all deaths to 1928. It also contains data describing many other life course events including records of arrival and departure, court appearances, military enlistment, property valuations for taxation purposes, details of bank accounts, census and muster returns, street directories and hospital and pauper admissions. As well as individual level data, the collection also contains additional tabulated data from census returns and statistical reports and digital images of many of the original records from which transcripts have been taken.

The dataset has resulted from collective research endeavours dating back to the 1990s and has been put together by researchers working at multiple institutions. Within Australia these include the Universities of Tasmania, Melbourne, Flinders, Monash, South Australia, Newcastle, New England and Griffiths. In addition, researchers at Guelph (Canada) and Liverpool and Oxford (UK) have supplied both data and expertise. The research collection has also benefitted from productive partnerships especially with the Female Convict Research Centre, the Tasmanian State Library and Archive, the Port Arthur Historic Site Management Authority as well as other collecting institutions. Data is continually added as a result of a volunteer transcription program organised through DIGIVOL, the crowd sourcing platform of the Atlas of Living Australia.

The Tasmanian Historical Dataset differs from other initiatives in a number of important respects. While historical life course and intergenerational record sets are now relatively common, it is unusual for these to contain data harvested from multiple sources (in this case drawn from more than 60 different archival series). The cosmopolitan nature of the collection reflects the diverse interests of the team of researchers who assembled it. They primarily work in the fields of economic and social history and historical archaeology, criminology and demography. The diversity of the collection presents some distinct advantages. The ability to interrogate multiple total count datasets in parallel is particularly powerful, providing opportunities to understand record keeping processes and other selection bias issues in greater detail than may be possible with more restrictive data collections.

Other differences reflect the peculiar nature of Tasmania's colonial past. A notable feature, for example, is the inclusion of life course data for 64,819 male and 13,673 female transported convicts. This includes a detailed record of all punishments inflicted on these individuals down to each day spent in a dark cell and each stroke of the lash. The public nature of the data is also unusual. As the dataset has been assembled as a result of a collaboration with local and family history researchers as well as archives and heritage sites it has for long had a life outside of the walls of academic institutions. It feeds information, for example, into the Tasmanian Archive search portal as well contributing to a number of educational and site interpretation tools.

This article starts with a brief description of Tasmania and its history. We then provide a more detailed description of the source materials, our approaches to record linkage and the structure of the datasets and the variables they contain. We conclude by making some observations on some of the problematics of digital record reconstruction using criminal and colonial archives as well briefly outlining some of the public aspects of the data and its potential research uses.

2 THE STUDY AREA

Tasmania, which used to be known as Van Diemen's Land, is the smallest state in Australia. According to the 2016 Australian census the island had a total of just 509,965 inhabitants — two-fifths of which were resident in the capital, Hobart. Despite its small size, Tasmania has a number of characteristics that make it of particular interest to researchers who wish to explore the extent to which the life experiences of individuals impact upon the health and socio-economic status of their descendants. Tasmania's relatively small population and defined geographical boundaries provide distinct record linkage opportunities. It is also a place with a long history of organised record collection.

The island was first occupied around 42,000 years BP. It was subsequently cut off from the Australian mainland by rising sea levels following the end of the last glacial maxima around 12,000 years ago. The next contact between Tasmanian Aboriginal peoples and the outside world occurred in 1642 when the Dutch East India Company commander, Abel Tasman, briefly landed on the east coast naming the

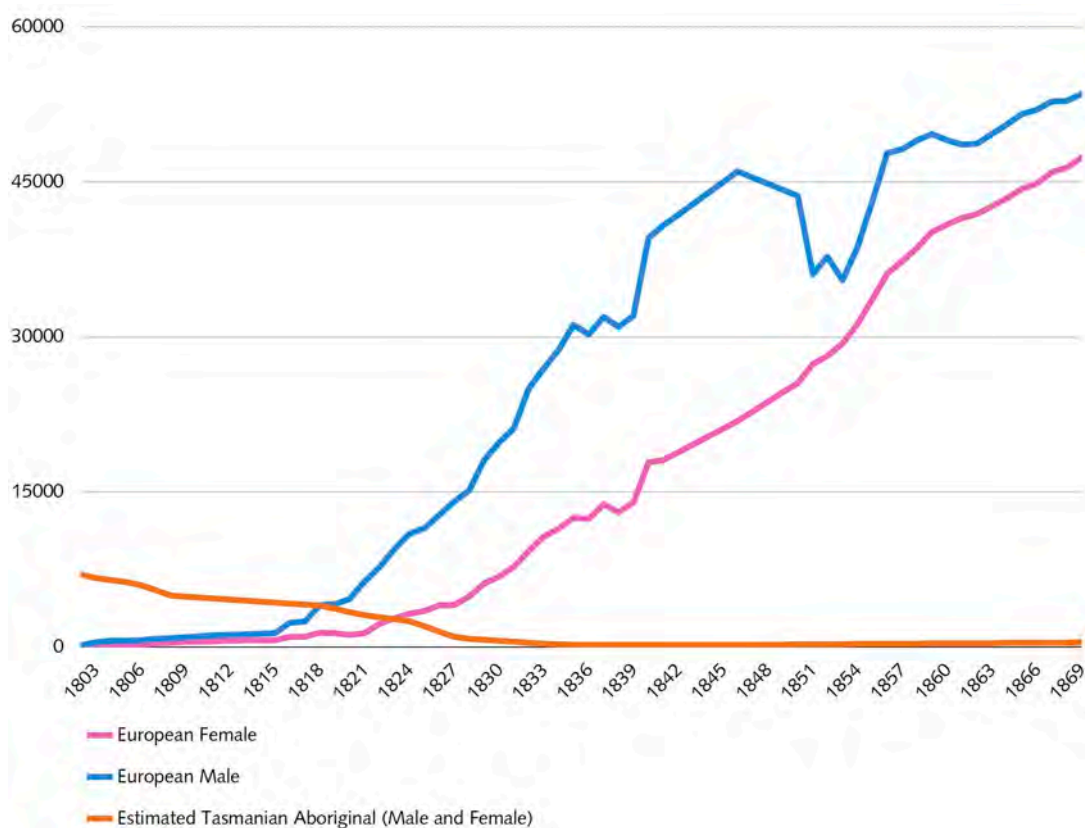
island Van Diemen's Land (it was renamed Tasmania in 1856). The first European settlement dates from the 1790s when sealing crews working out of Sydney occupied some of the offshore islands. Official settlement followed in 1803 when the British government sent landing parties to the island in order to secure the main anchorages in the north and south.

For the first fifty years following its colonisation Van Diemen's Land served as one of the principal penal colonies of the British Empire. During that time, it received at least 73,500 convicts — about 45% of all of those despatched to the Australian colonies. European unfree labour catalysed the process of colonisation. In effect the British used the labour of thieves to steal the island, an act of dispossession that culminated in frontier conflict and the enforced removal of the surviving Tasmanian Aboriginal population to offshore mission stations. As a result, the population of the island rapidly declined following first settlement (see Figure 1).

The convicts sent to Van Diemen's Land in the years 1803–1853 were predominantly tried in British and Irish courts, although small numbers arrived who were convicted in other British colonies including Mauritius, India, the Cape, New Zealand and the West Indies. They constituted the single most important source of colonial labour for the first five decades of the colony's existence. Rather than being kept locked behind forbidding institutional walls, the majority of serving prisoners were loaned or hired out to private sector masters. While some masters were former convicts, or the colonially born descendants of transported prisoners, many arrived free. A relatively small number of free migrants received a disproportionate share of grants of land as well as access to cheap convict labour. As the costs of maintaining a prisoner amounted to 59% of a free wage, this settler elite benefitted substantially from the offshoring of Britain's criminal justice system (Panza & Williamson, 2019).

The partnership with the private sector ensured that convict Van Diemen's Land shared much in common with colonial plantation economies. While private sector masters could not punish their unfree servants, they could bring a charge against them in a magistrate's court. These institutions were empowered to sentence a convict to undergo further punishment in a house of correction, a road or chain gang or a penal station. This ensured that two unfree labour systems, one run by the private sector and the other by the state, operated in parallel.

Figure 1 *Population of Tasmania 1803–1870*



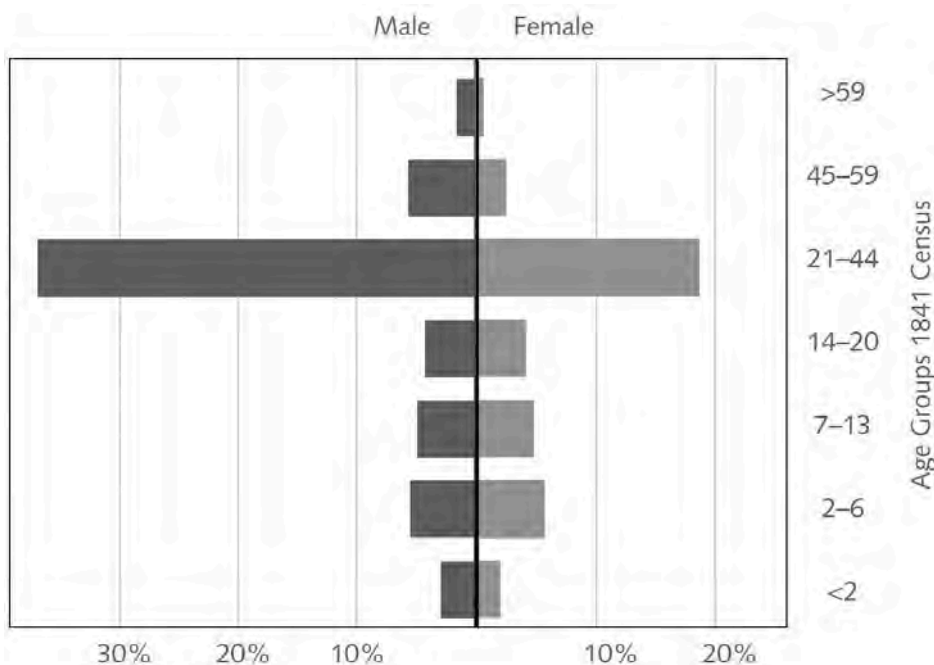
Sources: *Australian Bureau Statistics, 2014.*

From the 1830s a series of colonial initiatives attempted to attract alternative sources of free labour through assisted migration programs. These became particularly common after the cessation of convict transportation to the colony in 1853. These initiatives were particularly aimed at recruiting female migrants. The settler colonial population had a marked male skew as well as an age structure significantly different from that characteristic of Old-World populations (see Figures 1 and 2). The colony also experienced a marked temporary decrease in population (especially amongst males) as a result of the discovery of gold in the neighbouring colony of Victoria in 1851.

Perhaps because such a large proportion of 19th-century European migrants to Tasmania arrived as convicts, the colony developed a number of record-keeping systems which were unusual in both coverage and the detailed nature of their content. Although plans to introduce an annual census did not eventuate, 29 censuses were conducted in the period 1837–2016 — an average of one every six years. While it has long been the practice in Australia to destroy individual returns after the publication of each census report, the one exception is colonial Tasmania. Complete or partial returns are available for the six censuses conducted between 1837 and 1857. In addition, digitised tabulated data exists for all other censuses. These data were supplemented in the years before Tasmania joined the Australian Commonwealth in 1901 by annual statistical reports forwarded to London as part of the trans-imperial Blue Book system of colonial reporting.

Tasmania is also blessed with the second longest run of birth, death and marriage certificates in the Anglophone world — civil registration was introduced in 1838 (one year after England and Wales). These sources are available in digitised form from 1838–1899 and from 1970 to present. Plans are currently in train to digitise the remaining hard copy death certificates from 1900–1969. Numerous digitised parish records also exist. These are especially important for the years prior to the introduction of civil registration, but can also be useful after 1838. Some provide additional information such as the ship of arrival to the colony for example. Other series can be used to check the extent of under registration by particular religious denominations. Catholics, for example, did not always comply with civil registration in the belief that registration with the Church of Rome was sufficient.

Figure 2 Age and Sex Structure 1841 Census



Source: ABSTRACT of the Returns of the POPULATION and HOUSES in the different POLICE DISTRICTS, as defined in the Government Notice of the 27th September, 1841, Hobart, 1842. Historical Census and Colonial Data Archive. Retrieved from <https://dataverse.ada.edu.au/dataset.xhtml?persistentId=doi:10.26193/MP6WRS>.

3 CONTENT OF THE DATABASE

Data tables have been sourced from over 60 different archival series. For ease of reference these have been subdivided into the following datasets: 1) Convict; 2) Colonial Courts and Criminal Justice; 3) Census and Musters; 4) Departures and Arrivals; 5) Births, Death and Marriages; 6) Property and Financial Records; 7) Hospital and Pauper Admissions, 8) Military Records; 9) Maps and Plans and 10) Meteorological Data. The contents of these are listed in appendix 1 with the exception of 3.9 (Maps and Plans).

3.1 CONVICT

The men and women transported as convicts to Australia are of interest to historians because of the detailed way in which they were described. This process started with arrest and prosecution in Britain and Ireland and other trans-imperial courts. Links between records held in the Tasmanian Historical Dataset and British criminal justice system records are contained within the Digital Panopticon website (<https://www.digitalpanopticon.org>). This is especially the case for those tried in London's Old Bailey.

Post-conviction convicts were held in British and Irish institutions before they were embarked on transport vessels bound for Australia. Archival series have been transcribed for several of these holding institutions including Grangegorman female penitentiary, Dublin, 1840–1852; Millbank convict prison, London, 1837–1846; Pentonville penitentiary London, 1842–1847 and all British Hulks (prison ships used to accommodate prisoners sentenced to transportation), 1837–1845. Many entries in these series are for individuals who were sentenced to transportation, but instead served out their sentences in Britain or Ireland or were pardoned. Others are for convicts who were sent to penal colonies other than Van Diemen's Land. These data are useful in that they can be used to explore the manner in which convicts were selected into various transportation streams. They also enable a detailed comparison of institutional death rates. Finally, many of the inmates in penitentiaries and hulks were interviewed and described in ways that were similar to the processes that accompanied disembarkation in the Australian colonies.

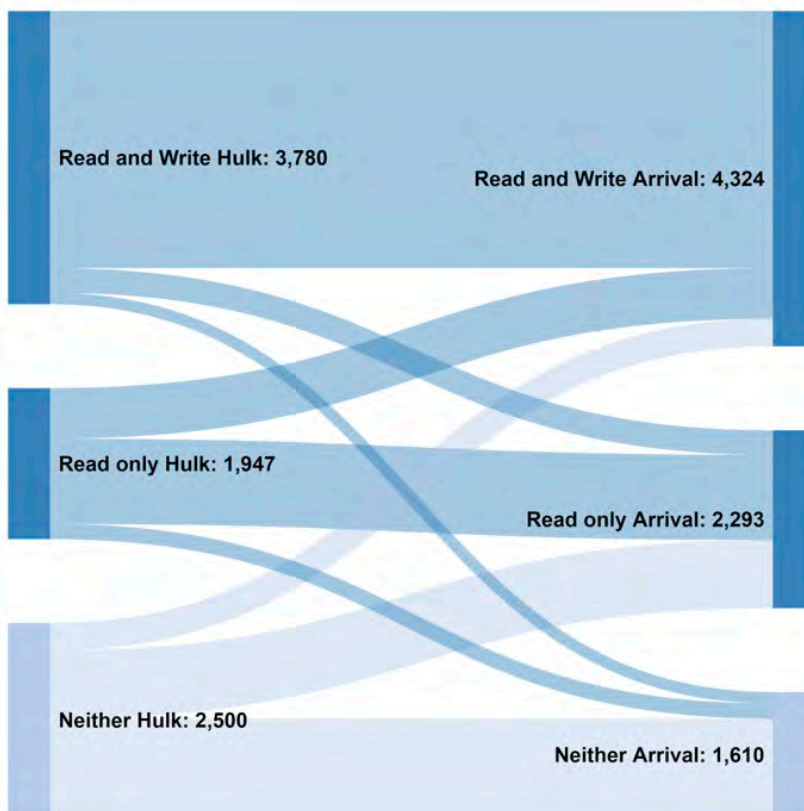
The ability to gather multiple observations for the same individual is useful as it enables an analysis of different data collection processes, as well as differences in response to similar questions over time. An example of this is provided in Figure 3 which explores differences in answer to questions about literacy levels provided by 8,227 male convicts on admission to the hulks in England and on arrival in Van Diemen's Land approximately eight months later. A notable feature of this Sankey chart is that more convicts claimed improvements in literacy levels than those reporting a decline.

The surgeon superintendent appointed to maintain health and discipline on each convict vessel was instructed to keep a record of all treatments administered during the four-month voyage to Australia. The journals for 289 voyages to Van Diemen's Land sailing in the period 1817–1853 were imaged in the National Archive, Kew, London and the treatment lists contained in each transcribed.

On landing each convict was provided with a police number — an early example of the use of identifiers to aid the tracking of individuals. This number, together with the details of the ship of arrival, the length of their sentence and place and date of trial, were used to index all records subsequently generated in the colony. Every convict was also interviewed prior to disembarkation. Together with a record of their past interactions with the court system forwarded on the same vessel that carried them into exile, these testimonies provide details of place of birth, age, next of kin, religion, occupation, level of literacy and a statement of previous life circumstances including the number of times each man and woman had been convicted. Each new initiate into the penal colony was also measured to the nearest quarter inch and described. This process included the documentation of scars, injuries and other deformities, as well as eye and hair colour and the documentation of any tattoos.

Many additional archival series were generated in order to assist with the operation of this complex unfree labour system. The most important of these were the conduct records (also known as the black books). This elaborate series of registers contain summaries of every colonial court encounter by a convict still under sentence, as well as many charges brought against former convicts long after they had become free. These include a detailed enumeration of every punishment down to each stroke of the lash applied to a convict's back, each day spent in solitary confinement or at hard labour.

Figure 3 Differences in response to questions about literacy provided by male convicts on admission to British hulks and subsequently on arrival in Van Diemen's Land (n=8,227)



Sources: Tasmanian Archive Con 13, 18, 33 and 77. The National Archive, HO 9 Prison Hulk Registers and Letter Books, Euryalus Register 1822–1836, 1837–1843; Fortitude Register 1837–1843; Ganymede Register 1837–1845; Justitia Register 1837–1844; Leviathan Register 1837–1844; Warrior Register 1837–1845; York Register 1837–1845.

Figure 4 Conduct record for Sarah Anne Adamson, Police Number 193

193	Adamson Sarah Anne		Transported for Larceny from the person Gaol report before not guilty House of correction often plunders Stated offence stealing 5/- from the person once for drunk once for drunk discharged Single child father Ralph Watson dead - Subsequent Reports - Indebtedness Well behaved but rather noisy with her tongue														
Tried	G.B. Court. 12 th May 1845. 10		Trade.	Height.	Age.	Complex.	Head.	Hair.	Whiskers.	Visage.	Forehead.	Eyebrows.	Eyes.	Nose.	Mouth.	Chin.	Native Place.
Embarked			At. Servant.	5/23/28.	28.	Brown	Says	Brown		Round	M.H.	B. Brown	Blue	Long	Small	Small	Worcester
Arrived	7 th Nov. 1845. Protestant		Marks Scar over left eye R O S B on left hand														
Period of Gang Probation			Six Months														
Station of Gang			Anderson														
Class			B 15/5/26 3 rd 15.0.40														
Offences & Sentences.																	Remarks.
6 th March 47 for 16/- absent without leave & not behaved																	
Labor still faulty 11.11.47. Delivered up on illegitimate grounds of 2 years 7 months 10 days 10 months																	
June 1847 Aug 1/2 1848 (Lovers) Absent without leave 2 Months, had labor (L. & M.) Sept 21/1848 Dec. 1848																	

Sources: Tasmanian Archive Con 41-1-7, image 177.

Other digitised records include notices of appointments to the colonial police (the latter was largely staffed by serving convicts), details of prisoners transferred to different private sector employers, descriptions of prisoners who had absconded, applications for permission to marry and information about the receipt of tickets of leave (an early form of probation) and the issue of pardons and certificates of freedom. Before 1840 convicts in private sector employment were not supposed to be paid a wage. After that date a system of payments was introduced in order to distance penal transportation from any association with slavery. Such payments were tracked through a series of registers that specified the duration of each passholder contract and the amount the convict would be paid. Finally, the deaths of convicts who were still under sentence were not entered into the civil registration system. Instead these were recorded in a separate series of convict death registers.

3.2 COLONIAL COURTS AND CRIMINAL JUSTICE RECORDS

Three levels of courts operated in Tasmania: magistrates' benches and police courts (also known as lower courts and or Petty Sessions); quarter sessions and the supreme court. Records for all defendants, charges and verdicts in Tasmanian Supreme Court cases 1824–1939 were collected as a result of the Prosecution Project and are available through Australian Historical Criminal Justice Data Dataverse administered by Griffith University.¹ The Quarter Session records are yet to be imaged and transcribed. Selected lower court records have been converted to machine readable format. These include a series for the Coal Mines, a punishment site located on the Tasman Peninsula and the Record of Cases heard against women in the Hobart Petty Sessions 1846–1854 (Tasmanian Archive, LC251/1/1 and 2).

After 1865 details of both convicted and discharged prisoners were routinely published in the *Tasmanian Police Gazette*. Data has been collected from 50,387 of these notices covering both male and female prisoners in the period to 1924. Available information includes date and place of conviction, charge and sentence, place of birth, ship of arrival to the colony in the case of those not born in Tasmania, age, occupation and physical description including height to the nearest quarter inch.

Figure 5 *Tasmanian Police Gazette list of discharged prisoners, 24 March 1876*

PRISONERS discharged from H. M. GAOLS and HOUSES OF CORRECTION, Hobart Town and Launceston, during the Week ending 22 March, 1876; Country Districts for the Week ending 18 March, 1876.

Name.	Ship.	Where Tried.	When.	Offence.	Sen- tence.	Native Place.	Age.	Height	Hair.	Remarks.
<i>Hobart Town.</i>										
Blakey, Robert	Equestrian 3	Hobart	15 Dec. '75	Idle and disorderly	3 mths	Yorkshire	60	5 8½	Greyish	F.S. Walks lame.
Doughney, Daniel, or. Dogherty	Offley	Ditto	20 Dec. '75	Ditto	Ditto	Isle of Man	57	6 0	Brown	F.C. Mole centre of forehead.
Upton, James	Palmyra	Hamilton	7 Mar. '76	Breach M. & S. Act	14 dys.	Cheshire	45	5 9	Dark brown	F.S. Little finger left crooked.
Mansfield, Wm.	P.Bomanjee 2	Bellerive	15 Mar. '76	Idle and disorderly	7 days	England	57	5 5½	Ditto	C.P. Scar on forehead above right eye.
Madden, Henry	..	Hobart	21 Mar. '76	Assault	3 mths	Tasmania	20	5 5½	Ditto	Free. Index finger (right) injured. Fine paid.
Harbuckle, Man- son	..	Ditto	21 Feb. '76	Ditto	1 mth	Ditto	18	5 1	Dark brown	Free.
Fox, William	Equestrian 3	S. C. Launceston	19 Mar. '72	On premises for an unlawful purpose	5 years	Derbyshire	67	5 7½	Grey	F.S. Mole under right eye.
<i>Launceston.</i>										
Davis, William	Moffatt 1	Longford	20 Sept. '75	Larceny	6 mths	Lincolnshire	60	5 6	Grey	F.S.
Golding, Daniel	Cambridge-shire	Launceston	24 Sept. '75	Ditto	Ditto	New York	27	5 6½	Brown	Free. Two women a bracelet and shield on left arm.
Brown, John	Derwent	Campbell Town	21 Feb. '76	Ditto	1 mth	Belfast	37	5 8½	Ditto	Free. A bracelet on left wrist.
Sullivan, Michael	Gazelle	Deloraine	Ditto	Idle and disorderly	Ditto	Limerick	51	5 8	Dark	F.S. Blind right eye, scar tip of nose.
Smith, Mary Ann	Margaret	Longford	20 Sept. '75	Larceny	6 mths	Dublin	40	5 4	Brown to grey	Free.
<i>Green Ponds.</i>										
O'Brien, John	Ld. Auckland 2	Kempton	10 Mar. '76	Ditto	7 days	Ireland	53	5 4	Brown	F.S.
<i>Torquay.</i>										
Stone, Joseph	..	Torquay	8 Mar. '76	Ditto	Ditto	Tasmania	18	5 9	Ditto	Free.

Office of Inspector of Police, Hobart Town, 24th March, 1876. JOHN SWAN, Inspector of Police.

1 https://dataverse.ada.edu.au/dataverse/australian_historical_criminal_justice_data

3.3 CENSUS AND MUSTERS

Although the convict population was mustered annually returns survive only for the years 1822, 1823, 1825, 1830, 1832, 1833, 1835, 1841, 1846 and 1849. The four musters conducted between 1830 and 1835 have been transcribed. Data includes information on the current place of employment of each convict as well as their police number and ship of arrival to the colony.

Returns for 14,870 households censused between 1837 and 1857 survive in manuscript form (see table 1). Although no return for an individual year is complete, the surviving returns are organised by parish. Even in years where only a small fraction of the original returns survive, these represent complete parish returns. These records have been digitised providing information on the name of the head of household, age, sex, religion and civil status (convict or free) of all occupants and the address, size and nature of the dwelling. All tabulated census returns contained in the original 29 census reports published between 1837 and 2016 are also included in the Tasmanian Historical dataset alongside a series of shape files which map changes in census collection districts over time.

Table 1 *Surviving household manuscripts Tasmanian census*

Census	Surviving manuscript returns	Estimated per cent of original
1837	1295	23.96
1838	124	2.17
1842	3895	53.40
1843	2562	31.74
1848	5136	53.96
1851	1740	15.47
1857	118	0.83
	14,870	

Source: *Historical Census and Colonial Data Archive*. Retrieved from <https://dataverse.ada.edu.au/dataset.xhtml?persistentId=doi:10.26193/MP6WRS>.

3.4 FREE DEPARTURES AND ARRIVALS

Information about free arrivals were sourced from the Tasmanian Archive index to free arrivals to Van Diemen's Land in the years to 1856. The file contains details of 24,232 arriving passengers. A companion file contains information on 114,452 departures in the period 1817–1858. Usefully this provides information on the ship of arrival in the colony (a great help in identifying former convicts). In addition, more detailed records are available for 10,631 assisted migrants arriving in the period 1852–1858 (Tasmanian Archive, CB7-1-13-20). These records contain information about age, sex, marital status, religion, native place, literacy, occupation, employer's name and agreed wage rate.

3.5 BIRTHS, DEATHS AND MARRIAGES

A total of 195,000 births were registered in Tasmania between 1838 and 1899 and 51,000 marriages, all of which have been digitally transcribed. A longer run of 155,000 digitised civil registered deaths for the years 1838 to 1928 is also available. Information includes age at death and cause of death. There are slight variations in the information included in birth, death and marriage certificates over time. Usefully, details of place of birth were included on Hobart death certificates from 1857, Launceston from 1886 and all death certificates from 1895. Information relating to surviving children is also included on 20th-century certificates as well as details of marriages and spouses. The transcriptions for all three series have been linked to digital images of the original records.

In addition to civil registered births, deaths and marriages, the collection also contains transcripts taken from ecclesiastical registers. These contain information on 19,723 baptisms and 8,828 burials many of which predate the introduction of civil registration in 1838.

3.6 PROPERTY AND FINANCIAL RECORDS

Many convicts arrived in Australia with goods and cash. These were held in trust by the colonial state while the convicts served their sentence. From 1829 on cash sums were entered into a Convict Savings Bank managed by the directors of the Derwent Bank. Between the years 1845–1863 the Hobart Savings Bank kept a record of all customers who had opened bank accounts, both free and unfree. This dataset consists of 12,240 records and includes information on age, sex, occupation, place of residence, civil status (convict or free) alongside a physical description, including height — information that was committed to file in an attempt to stop fraudulent access to accounts (Tasmanian Archive, TAHO NS1167).

Several other record series in the dataset include information about place of residence. Thirty-five trade and street directories for Hobart and Launceston were published between 1825–1854. These contain details of 26,000 addresses, many relating to shops and businesses. Tasmania was unique amongst Australian colonies in that it published annual valuations of all properties. These commenced in 1847 in Hobart, were extended to Launceston in 1853 and the entire colony in 1858. Entries include information on the occupier, the owner, the nature of the property and its annual rateable value. The dataset currently contains a complete run of Hobart Valuations for the years 1853–1883 and a total count of properties in the colony in the years 1858 and 1861, a total of 153,291 entries.

3.7 HOSPITAL AND PAUPER ADMISSIONS

The dataset contains multiple hospital, invalid and pauper admissions and discharges including: 2,650 recorded deaths in the Hobart Hospital (1864–1884); 1,871 admissions to St Mary's Hospital Hobart (1853–1862); 4,488 admissions to the New Norfolk Psychiatric Hospital 1830–1901; and 35,161 pauper admissions and discharges covering the period 1858–1952.

3.8 MILITARY RECORDS

Data for 15,234 Tasmanian-born soldiers and nurses who enlisted in WWI has been harvested from attestation papers held in the Australian National Archive. This information includes date of birth, name and address of next of kin, height to nearest quarter inch, weight in stone and pounds and both expanded and unexpanded chest measurement.

3.9 MAPS AND PLANS

Tasmania has a rich resource of cartographic and planometric material. The collections are predominantly held in two main institutions: the Tasmanian Archives and Land Tasmania. The former holds over 180,000 records, of which 26,000 are digitised. Its collection encompasses exploration charts, road and town charts, architectural plans and elevations, drawn from a range of government, institutional and private sources. Land Tasmania retains title and deed plans relating to the administration and sale of land back to the early 19th century. Links to both collections are contained within many digitised records held within the Tasmanian Historical Dataset.

3.10 METEOROLOGICAL DATA

Commencing in 1825 daily weather measurements for Hobart were routinely published in the press. Information varies in detail from year to year but always includes minimum and maximum daily temperatures and barometer readings as well as wind direction. There are some gaps in the series notably between February 1827 and April 1838. Weekly averages are available for some years where daily data is missing.

4 DATA CODING

Our approach to data coding has been guided by the Intermediate Data System ([Alter & Mandemakers, 2014](#)). All data has been separated into two types of entity: persons and contexts. Each of these may be assigned attributes such as a person's sex, age, occupation and civil status at any given point in time, or specific events such as court appearances, marriages, admissions and arrivals, departures

and discharges. Contexts are usually locations and can be nested one within the other. Thus, *Spring Grove* is a property in the district of Patterson's Plains that lies within the parish of Selby which is itself contained within the Tasmanian county of Dorset. Individuals may be linked together by familial or social relations. They might also be linked to particular contexts at any point in time. Thus, several individuals might reside at the same property or be barracked in the same building. Individuals can also share the same place of employment. All individuals and contexts are assigned unique identifiers. All attributes and relationships are assigned codes managed through coding dictionaries. These list each occurrence of every variable and the codes that have been assigned to it. Wherever possible international coding systems have been used to populate these dictionaries. Inevitably, however, the processes of analysis have involved some adaptation or the creation of new codes. These variations have been documented within each dictionary.

4.1 INFORMATION ABOUT OCCUPATION

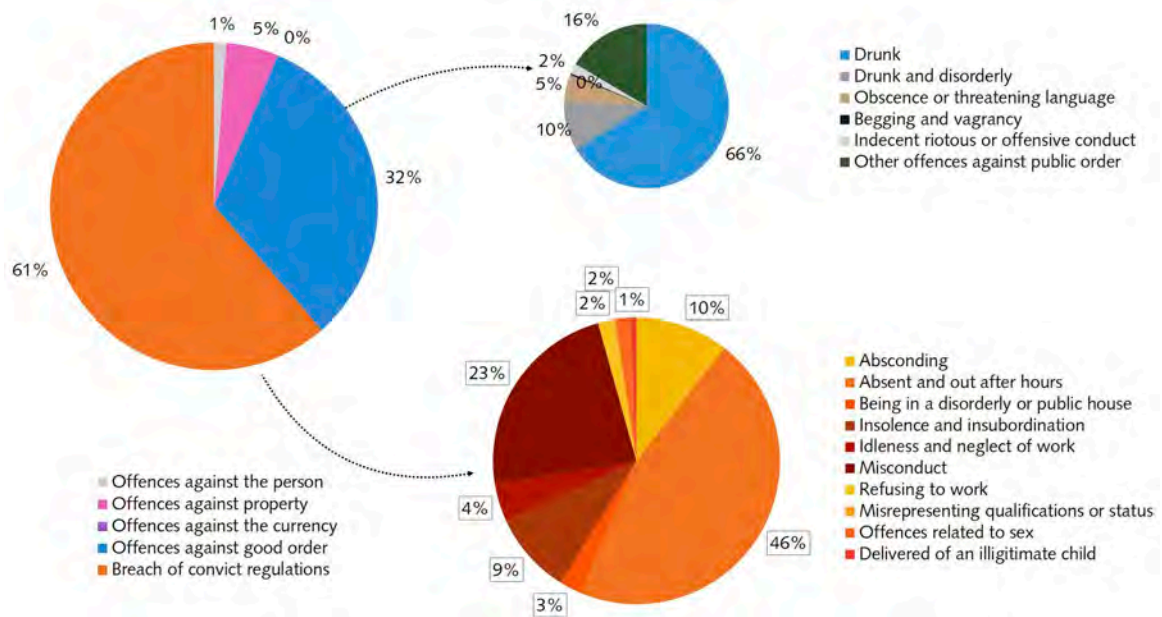
All occupational information has been coded according to HISCO, the *Historical International Standard Classification of Occupation* (van Leeuwen, Maas, & Miles, 2002). This fine-grained system of organising occupational descriptions according to the nature of each task has been mapped onto HISCLASS, a related coding that uses information about occupation to proxy social stratification (van Leeuwen & Maas, 2011). The HISCLASS handling of some occupational groups sits oddly with what we know about New-World social structures. This is particularly the case with agricultural and pastoral landholders. In order to take account of these differences, as well as to compare results with previous studies, other industrial and social stratification codes have been added. These include an industrial classification based on the British 19th-century census first applied to convict data by Lloyd Robson (1965) and Armstrong's social classification system and the Nicholas and Shergold variant of this (Nicholas, 1988). As new occupational data is collected, the dictionary is used to automatically code all descriptions of job titles and work processes previously encountered, ensuring consistency of classification across different datasets.

4.2 VOYAGE OF ARRIVAL

Identifying the means by which individuals arrived in the colony can be important for record linkage. It was also an important determinant of social status. Identifying a particular voyage of arrival is often complicated by vessel naming practices. Multiple ships named the *Asia* sailed to Tasmania for example. In addition, many vessels made more than one trip to the colony. The assigning of unique voyage codes for convict vessels has proved particularly important in record linkage. It is also important in that the same coded attribute can be used to identify individuals who share a relationship as 'shipmates'. At time of writing work is progressing on a related project to identify free migrant voyages.

4.3 CRIMINAL JUSTICE DATA

There is no agreed international coding system for classifying historical information about crime although several contemporary classification systems exist which can be mapped onto one another. This includes the International Classification of Crime for Statistical Purposes (United Nations Office of Drugs and Crime, 2015). A particular problem with matching 19th-century Tasmanian criminal justice data to these schema is that many of the charges brought against serving convicts are not commonly encountered in contemporary judicial systems. This reflects the manner in which the labour of prisoners was outsourced to the private sector. As a result, magistrates' courts often heard offences that revolved around the non-performance of work or the perceived degree of effort or diligence that convict workers were said to have displayed by masters or others charged with supervising them. Many were prosecuted for 'refusing to work' and malingering. 'Concealing a pregnancy' was even regarded as an offence — convict women who fell pregnant were routinely sent to the house of correction for punishment. Others were charged with 'insolence' or infractions of the rules governing the management of different institutions. To provide an illustration of the extent of the problem, while 67,606 magistrates' bench charges are recorded in the conduct records for 13,415 convict women, only 1% of these involve offences against the person and only 5% were for offences against property. By contrast, 61% were for breaches of the rules and regulations governing the conduct of prisoners under sentence (see Figure 6). As convict administrators had to account for the distribution of such charges, they developed a classification system which we have utilised to code this data. As with information about occupation, we have created coding dictionaries to ensure standardisation of coding across information retrieved from multiple archival series.

Figure 6 *Classification of charges brought against convict women*

Note: Number of women = 13,415; Number of recorded court appearances = 67,606; Number of charges = 84,344.

Sources: Tasmanian Archive, Con 40 and Con 41 series.

4.4 INFORMATION ABOUT MORBIDITY AND MORTALITY

Nineteenth-century causes of death and diagnoses can be difficult to map onto contemporary classification systems. To address this we adopted the 32-category system created by Rebecca Kippen, which usefully combines aspects of William Farr's 19th-century nosology with the contemporary international classification of diseases. This system is sufficiently broad to be analytically meaningful while at the same time specific enough to enable particular mortality and diagnostic trends to be plotted over time (Kippen, 2011). While Kippen's original schema was developed to code causes of death, our datasets also include information about other episodes of ill-health. In order to capture data about some events that were commonly diagnosed but rarely resulted in death, we have included some additional categories (see Table 2)

4.5 GEOLOCATING DATA

Many contextual variables have been geolocated including place of birth, conviction, incarceration and work. A Geographic Information System (GIS) has been used to geolocate historic maps and plans to modern survey and archaeological information. From this we have been able to digitise data at multiple scales: historic parishes; hundreds; counties; local government areas; townships; streets; buildings and even rooms. Nested spatial 'containers' or shape files are created as part of this process. Non-spatial information can be linked to these digital contexts effectively populating spaces with people, processes and products.

Granularity is an issue commonly encountered in historical research. This problem arises when information about place is recorded in ways that enable a more precise identification of location in some cases compared to others. Thus, while some convicts provided the name of the street they were born in, it was more common to report a parish of birth. Others still only provided information about their county of birth—a particularly common occurrence with convicts from Ireland. In these instances, we used the code for the county town but created an additional variable to inform users that this was a proxy location and that the precise place of birth was unknown.

Table 2 *Classifying information about cause of death and disease*

Code	Category
1	Accident
2	Convulsions and teething
3	Debility and marasmus
4	Diarrhoea and dysentery diseases of the blood and blood forming organs
5	Diseases of the circulatory system
6	Diseases of the digestive system
7	Diseases of the eye and ear
8	Diseases of the genitourinary system
9	Diseases of the musculoskeletal system
10	Diseases of the nervous system
11	Diseases of the respiratory system
12	Diseases of the skin and subcutaneous tissue
13	Endocrine, deficiency and metabolic disorders
14	Influenza
15	Malingering
16	Measles
17	Mental and behavioural disorders
18	Nausea
19	Neoplasm
20	Old age and decay
21	Other fever
22	Other infectious diseases
23	Other tuberculosis
24	Paralysis
25	Parasitic disease
26	Pregnancy, childbirth and the puerperium
27	Respiratory tuberculosis
28	Scarlet fever
29	Whooping cough
30	Sexually transmitted diseases
31	Suicide
32	Unclassifiable
33	Unknown
34	Unspecified natural causes

Source: (Kippen (2011) and Maxwell-Stewart and Kippen (2015)).

A related issue is the difficulty of locating names shared by more than one place. There are several places that are named Newcastle in the United Kingdom for example. While these are commonly distinguished by reference to local geographical features such as Newcastle-Under-Lynne and Newcastle-Upon-Tyne, this is not always specified in the original record. In terms of convict places of birth that might refer to multiple locations, we adopted the practice of geolocating to the location closest to the court in which the convict was sentenced to transportation.

For places within Tasmania we have adopted similar measures to cope with granularity by adding a resolution variable using three values. For the geocoding of places of incarceration or work, those which could be precisely mapped were marked as being geolocated with a 'high' level of resolution (Tuffin & Gibbs, 2020). Places which were located from poorly-geolocated maps, or were locatable to an area only (such as a precinct or parish), were marked as 'medium'. Other locations which could not be pinpointed to a specific location or area were mapped to townships or districts and accorded a 'low' level of resolution. The list of geolocations, along with their site-specific codes, has been archived within the Australian Gazetteer of Historical Place Names.

We have used geolocated historic maps and plans, survey and archaeological data to facilitate the spatio-temporal reconstruction of some built landscapes. Thus, a former penal station, Port Arthur (1830–1877), has been dynamically mapped across its 47-year period of operation (Tuffin et al., 2019). This has enabled the digital reconstruction of buildings, individual spaces, walls, fences, roads and workplaces. Where these have been named in the charges brought against convicts serving at this penal station, it has proved possible to link this data to each digitised context allowing offences to be mapped in time and space.

4.6 MAPPING CHANGES IN CENSUS DISTRICT

Shape files can play a particularly important role in linking individual level data to tabulated census returns. While each shape file can be tied to a census table, in order to analyse regional change between censuses it is necessary to identify differences in regional collection boundaries between censuses. The last census, conducted in 2016, was organised around a system of mesh blocks. These are the smallest geographical area defined by the Australian Bureau of Statistics and are primarily organised around land use. It is thus unusual for a mesh block that contains primary production land to include areas zoned as commercial, residential or parks. Each mesh block has been designed to be large enough to protect against accidental disclosure of confidential information. To this end, the majority of populated mesh blocks contain between 30 to 60 dwellings. Many 2016 mesh blocks, however, are completely unpopulated.

We have retrospectively mapped the 2016 mesh blocks onto previous census collection districts. Over the course of this exercise we have identified three types of alignment issue. The first of these can be attributed to differences in mapping standard. Nineteenth and twentieth century maps that depict the boundaries of parishes and local government areas were not surveyed to current standard. As a result, it often appears that boundaries do not align, although examination of underlying topographical features reveal that this is entirely due to mapping inconsistencies. Where we have encountered such 'cadastral noise', we have used the 2016 mesh blocks to redraw historic boundaries.

The second type of misalignment is caused by land use patterns. On occasions a current mesh block includes areas on both sides of an historic boundary, but on closer examination one part of the area bisected by the historical division contains settlement and the other does not. The unpopulated section of the mesh block typically consists of pastoral land. Where this has occurred, we have again used the 2016 boundaries to redraw the historic boundaries.

While in most cases it has proved possible to match mesh block boundaries to historic parishes, local government areas and subsequent census districts, there are occasions where boundaries have been redrawn in ways that problematise comparisons between successive tabulated census returns. In these cases, we have used the 2016 mesh block boundaries to highlight where these major realignments have occurred and have provided sufficient documentation to alert subsequent users.

5 LINKAGE BETWEEN SOURCES

Our general approach is to internally link each record series before linking across series. Thus, the valuation rolls are internally linked so that successive annual records which record the same occupier in the same address are formed into chains. Likewise, births are first linked to parental marriages in order to link the same individual in both records and define familial relationships. Both lists are then matched in a subsequent exercise aimed at locating individual and family records within households.

Linkage is an iterative process. We first generate standardised lists of key variables especially first name, surname name and ship of arrival. Automated record linkage queries are then run using these cleansed variables. All matches are then checked and cleaned using a duplicate query. Soundex codes are used in subsequent iterations. Linkage weights are employed to evaluate each step of the process and these are retained within datasets in order to assist with subsequent iterations. Finally, remaining unmatched records are examined by hand.

Other record linkage processes are used to match owners to businesses, businesses to places of residence and magistrates to police districts. Some record series also contain attributes that ease the process of record linkage. WWI attestation papers, for example, provide detailed information of next of kin as well as recording age in years and months. This considerably facilitates linkage with birth certificates (Inwood, Kippen, Maxwell-Stewart, & Steckel, 2020). Similarly, criminal justice series routinely contain attributes that lend them to linkage. This includes information about former dates of conviction and sentences. In the case of convicts transported to the Australian colonies, other identifiers were used to help administrators retrieve records pertaining to the same individual filed in different registers or correspondence series. These include police numbers and the name of the ship that transported each convict into exile. As a result, it is possible to link the records for serving convicts with a great deal of certainty (Maxwell-Stewart, 2016).

Record linkage for time-served convicts is more problematic. After all, the men and women 'lagged' to Australia had a vested interest in escaping their past. Name changes were commonplace as former convicts tried to reinvent themselves. The task of tracking emancipated prisoners post-release is more challenging for men than women, a reversal of the normal paradigm. Any convict who wished to marry while still under sentence had to apply for state permission. A much higher proportion of convict women compared to men are named in the registers that governed these processes, a reflexion of the colonial sex imbalance. Since it is relatively straightforward to locate a marriage certificate if both bride and grooms' names are known, the rate of linkage for convict women to marriage certificates is high. Post-sentence migration rates for former convict women are also lower than for men, a smaller proportion left for Victoria following the discovery of gold in 1851 for example (see Figure 1). As they were less mobile and more often embedded in family structures, they are more visible than might be expected although many women in de facto relationships changed their name to their partners name adding a further level of complication.

Despite these difficulties, many colonial record-keeping systems contain clues to identity. One of the reasons convicts sought to hide the details of their former lives is that the manner of arrival in the colony was regarded as a marker of status. Both state- and church-administered record systems sought to include identifiers that could help track legal status. It was common for individuals to be described as 'native' — that is colonially born — or came free, an indication that they had arrived in the colony as a migrant and not a prisoner. Former convicts on the other hand often had their records marked: C.P, standing for conditional pardon, or F.S., free by servitude, or F.C. free certificate — all indicators of former servile status. The ship that a person arrived in the colony on could also reveal much about former legal status. While such annotations are particularly common in criminal justice systems, they were also included on other records including church burial records, hospital admissions and registers of departures from the colony.

Some records also include descriptions of individuals. While this is more common with criminal justice records some bank account registers also contain physical descriptions. Eye and hair colour, height, and descriptions of scars, physical deformities and tattoos can provide useful identity pointers. While it is difficult to use this information to assist automated matching, it can provide a useful check for evaluating problematic matches. Physical descriptions can be particularly useful aids for linking records that could not otherwise be matched as a result of the use of aliases or other name changes. While the heavy reliance the convict system and subsequent colonial record-keeping systems placed on

descriptions as a means of checking an individual's status eases the issues of record linkage, it is dangerous to assume that those described in these records are representative of all former convicts. Just as convicts who marry are easier to trace in subsequent records, so are those who continued to have interactions with the criminal justice system.

6 IDENTIFYING SELECTION BIAS

Selection bias presents an ever-present challenge for historians. Since archival records were created in the past, researchers must make assumptions about the extent to which the resulting data is representative of the particular issues they wish to study. There are two aspects to this. First, the degree to which the information they utilise is representative of the wider record collections from which that material has been drawn; and second, the degree to which those collections reflect the historical realities that the researcher wishes to shed fresh light on. The use of total-count data can reduce the risk associated with the first of these processes, although surviving records may not be representative of the full range of information originally collected. Nevertheless, any attempt at digital reconstruction is likely to highlight gaps in a series and provide a means of estimating the extent of undercounting.

The digitisation of multiple series can help to explore the second type of selection issue. Comparisons of the ways in which individuals are described in multiple series can throw considerable light on original data collection processes. It can also identify individuals absent in one record but present in another. Both processes can yield information about the ways in which men and women were selected into different record-collection exercises. They can also be useful in reconstructing human agency. Most records are the product of an encounter between a person or household and the state or church. The way in which similar questions are answered in different contexts can reveal much about individual circumstances at the point in time when each record was formed (Maxwell-Stewart, 2016).

As others have argued, all archives were developed as administrative tools and as such are neither passive observers of the past or static entities. Because they were created with a particular purpose in mind, the way in which individuals are represented within archival collections reflect the concerns and prejudices of successive administrations. This is perhaps particularly the case with criminal justice and colonial records, series that were created in order to aid the policing and control of particular sub-populations. It is for this reason that Ann Laura Stoler argues that archival series need to be first read along the grain before an attempt is made to co-opt them for other research purposes (Stoler, 2009). A key rationale behind the Tasmanian Historical Dataset is that the assembly of digital record series in parallel will aid the kind of deep read advocated by Stoler — helping shed light, not only on selection processes, but also the underlying rationale that caused some individuals to be omitted from some series and described in particular ways in others.

7 USES OF THE TASMANIAN HISTORICAL DATASET

Reflecting its collaborative origins, the Tasmanian Historical Dataset has many different potential uses. In the following section we summarise some of the ways in which the information it contains has been employed to date as well as outlining future work.

7.1 LIFE COURSE AND INTERGENERATIONAL ANALYSIS

Multiple mechanisms are known to shape the health outcomes of both parents and their children (Kuzawa & Eisenberg, 2014). Maternal literacy is powerfully associated with improved intergenerational outcomes, for example, while elevated levels of alcohol consumption during pregnancy can stunt offspring growth and retard cognitive development (Riley, 2001; Rose & Cherpitel, 2011). The thrifty phenotype hypothesis posits that a history of maternal undernutrition may trigger foetal responses that favour the development of critical organs at the expense of others, leading to increased risk of chronic illness later in life (especially cardio-vascular disease and type-2 diabetes). Recent work also links thriftier

metabolism mechanisms in early life to poorer cognitive, immune system and reproductive development (Pike, 2016). Genetic factors play a role in the determinants of intergenerational health, although the difficulty of distinguishing genetic and environmental pathways is complicated by interactions between the two. There is evidence that effects associated with trauma exposure, for example, can be transmitted across generations via the chemical coating of chromosomes (Kellerman, 2013).

Intergenerational datasets composed of many linked life-course events are needed to explore these underlying causes of familial inequality. The ideal study population would consist of those who experienced a set of well-defined adverse circumstances and a suitable control group which escaped these particular experiences, but otherwise shared many characteristics. Such populations are rare. Gavrilova and Gavrilov (1999) find that most available data for the intergenerational study of longevity in Europe, North America and East Asia, have significant shortcomings. The data often target highly specific or localised populations and lack one or more important features, such as information on women, background information on socioeconomic conditions, the timing of stressful events for individuals, causes of death, individual characteristics, or family background. Their study highlights the need for new data to fill the evidence gaps and support a more convincing exploration of the transmission of inequality across generations. Tasmania is perhaps an exception to this general rule.

A particularly important feature of the Tasmanian Historical Dataset is that it includes life-course information for men and women that arrived as convicts and assisted migrants. Although these two populations migrated under different circumstances, similar detailed information is available for both. This includes details of place of birth, age, occupation and literacy. Thereafter the experience of the two groups differed markedly. While the convicts were subjected to punishments, including solitary confinement and hard labour, the assisted migrants were not. The abundance of civil registration data in Tasmania enable, not only the tracing of convict and assisted migrant cohorts to death, but also the identification of their Tasmanian-born children and grandchildren. An additional important attribute is that those children and grandchildren were raised in New World environments radically different from those that shaped the early life experiences of their parents.

Work to date suggests that some punishments, particularly solitary confinement exposure, cut short the life expectancy of convicts (Kippen & McCalman, 2015; McCalman & Kippen, 2020). While further work needs to be undertaken to determine whether these effects impacted upon the lives of subsequent generations, preliminary evidence provides an indication that the children of transported convicts experienced some benefits as a result of the misfortune of their parents. They were tall compared to Old World populations for example. Moreover, colonially born prisoners born to mothers who had been transported as convicts were taller than those whose mother had arrived free. This surprise finding probably reflects the smaller number of children born to emancipist mothers and hence the greater availability of resources per head in ex-convict households (Maxwell-Stewart, Inwood, & Stankovich, 2015).

As with the intergenerational transmission of health inequalities, there are multiple mechanisms that might explain the perpetuation of prosecution histories across generations. These include poor parenting, poverty of expectation and opportunity and the genetic transmission of conditions likely to heighten the risk of arrest (for example, schizophrenia). Such risks are often heightened by policing strategies. The children of those 'known suspects' are likely to be more severely policed than others. An island originally populated by transported convicts which is also blessed with long runs of digitised criminal justice and civil registration data, Tasmania provides an ideal opportunity to explore the different pathways by which prosecution risk might be transmitted from one generation to another (Godfrey, Inwood, & Maxwell-Stewart, 2018)

7.2 ENVIRONMENTAL DETERMINANTS OF WELL-BEING

Much of the literature pertaining to the mechanics of disadvantage, and its persistence through generations, has been formulated around social determinants. For example, there is a rich literature examining the links between education and future earnings and productivity, suggesting that these relationships may be intergenerational (Goldin & Katz, 2001). These pathways are complex. Thus, higher material well-being may lead to better educational outcomes not solely through the availability of financial resources, but through better health and higher life expectancy, which in turn generate greater incentives for investment in education. Far less attention has been given to how factors such as housing and the development of urban infrastructure feed into this process. This is somewhat

surprising given that the dwellings in which people live provide them with the most elemental of protections through shelter from environmental conditions and constitute a basic human need. It follows that dysfunction at this level will potentially have a profound impact over the development of human capital and subsequent life-course outcomes.

A related issue is the extent to which social mobility might exacerbate or mitigate early childhood disadvantage. Most measures of social mobility are based on occupational data (or in the case of 19th-century women, the occupations of their husbands as recorded on a marriage certificate). These measures are sensitive to age effects and changes in wage rates over time. They provide at best a snapshot at a particular point in life. The availability of annual housing valuation data provides an opportunity to create a more robust set of measures to explore these issues. It is certainly unusual to have a continuous measure which can be used to analyse the impacts of changing home ownership or variations in the relative value of an individual's place of residence over the life cycle. The ability we have to examine the effects of place and value of residence across generations is particularly rare.

7.3 SPATIAL ANALYSIS OF LABOUR RELATIONS

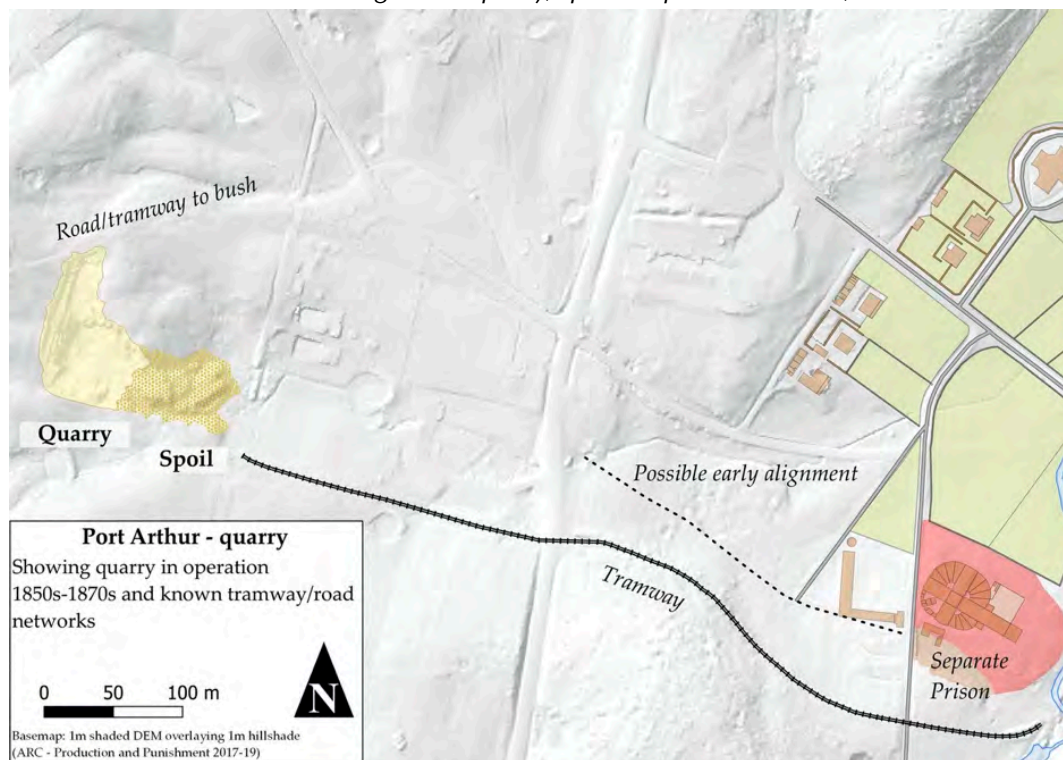
Many archival records contain references to locations that can be converted to multi-scalar and multi-temporal objects through the linking of historic and contemporary geographical data. Thus, notices of convict transfers between properties and the individual labour contracts signed by convicts in the period after 1844 can be geolocated in order to reconstruct the seasonal flow of labour between different sectors of the economy. Such analysis is useful in that it can provide an indication of the ways in which human capital — as measured by age, sex, skill and literacy rates — influenced contract length and individual levels of remuneration (Meredith & Oxley, 2005).

The charges recorded in a conduct record contain multiple references to place. These include the location of employment, the court where the case was heard and the site the convict was sent to undergo punishment. The specifics of each charge summary often contain additional geolocatable information such as the name of an inn or a particular building where an offence was said to have occurred. Other records can be used to determine the number of convicts at each employment location over time, enabling a reconstruction of prosecution risk across different types of urban and rural, and public and private workplaces. Analysis of these record collections confirms that those convicts with particularly valued skills were more likely to be provided the benefit of the doubt than was the case for those who were easier to replace. Punishment strategies also reflect the costs of maintaining a convict compared to hiring a free worker. When convict labour was relatively expensive, magistrate's benches were more likely to sentence prisoners serving in the private sector to stints of hard labour in a road party where they would be maintained by government (Maxwell-Stewart, 2015).

Similar forms of analysis can be used to explore the ways in which the deployment of convict labour impacted upon the landscape. Archaeological survey combined with the analysis of LiDAR (Light Detection and Ranging) remote sensing has been successfully employed to reveal the physical traces of convict labour (Tuffin, Roe, Gibbs, Clark, & Clark, 2020). Such sites include quarries, clay pits, sawpits, roadways, tramlines and buildings. Such contemporary landscape imaging can be aligned with historical maps and administrative records in order to answer questions about construction methodology, as well as provide greater insight into the extent and impacts of labour on both the environment and the individual (see Figure 7). Reconstituted punishment records can then be used to explore the extent to which levels of coercion varied across different work landscapes (Tuffin et al., 2019).

The ability to place convicts in individual workplaces can also aid analysis of worker agency and resistance. Using runaway notices placed in the *Hobart Town Gazette*, for example, it is possible to plot absconding patterns across different locations (see Figure 8). Such point-cluster maps illuminate regional variations in absconding and prosecution patterns over time (Tuffin & Gibbs, 2019). While many other convicts were punished for 'refusing to work', analysis of collective action has been hampered by the way in which this information was recorded on individual conduct records. The extent of collective punishment is only revealed when whole series are transcribed and linked by date and site of employment. This can also shed light on the ways in which different forms of action followed one upon the other. Convicts would often attempt to petition higher authority, before striking and then finally absconding when other attempts to address grievances had failed (Dunning & Maxwell-Stewart, 2002; Tuffin, Maxwell-Stewart, & Quinlan, 2020).

Figure 7 *LiDAR scan showing site of quarry, spoil heap and tramlines, Tasman Peninsula*



8 SUMMARY AND CLOSING OBSERVATIONS

This article has described the diverse set of records that make up the Tasmanian Historical Dataset. In many ways the collection reflects a 19th-century Antipodean fascination with classification. Tasmania's convict past ensured that it developed as a society where much information was recorded — a necessary part of keeping the unfree in line. Yet, concerns about the long-term impact of penal transportation ensured high levels of record keeping long after the last convict vessel had arrived. It was also a place, however, that excited a great deal of scientific interest — a product of the way its fauna and flora seemed unusual to European eyes. The Royal Society of Tasmania, founded in 1843, was a particularly enthusiastic promotor of the systematic collection of information. Such classification exercises were rapidly extended to include social data. The early introduction of civil registration and attempts to hold a census on a regular basis reflect this legacy.

The overlapping nature of the resultant record systems and relatively small population sizes have encouraged more recent attempts to use digital techniques to reconstitute and reanalyse these series. This has been a somewhat unusual enterprise in that it has involved a collaboration between archivists, researchers from a variety of different academic backgrounds as well as many family historians and local history groups. It has also differed from some other longitudinal historical exercises in the diversity of the record collections that have been assembled and linked.

This has some disadvantages. Navigating the collection can be confusing. Different records were generated under different circumstances, all of which require careful documentation. It also has strengths, however, in that it enables users to see how individuals were represented in multiple different datasets. In this kind of complex data environment, even an absence can be informative. The ability to bring multiple series into alignment will help users to better understand the selection processes that shaped the development of each record series.

Figure 8 *Locations convicts were advertised as absconding from, Van Diemen's Land 1839*

Source: *Hobart Town Gazette, 1839*

This in turn should lead to improvements in archival catalogues, finding aids and interfaces. There is a technical difference between a record and a document, in that a record has a known context and transactional history (Sternfield, 2011). The creation of archival networks can transform a loose assembly of documents into a collection of records. In a digital research environment, this is not a static process. Rather than disassociating information from its archival context, each systematic interrogation of a digital archive can add to a record's transactional history. Thus, cross-interrogation of similar series can assist in reconstructing the sequence of individual record production and in turn aid an understanding of both content and context.

Just as data creation has shaped so much of Tasmania's settler past, so the digitisation of that data is helping to shape its future. Once a mark of shame, Tasmania's convict era is now seen as something of a drawcard. Former penal stations and houses of correction have been repurposed as heritage sites and 19th-century houses built by convict labour have been turned into boutique accommodation. The ability to visualise the ways in which tens of thousands of life courses interacted, plot the route that convict voyages took, chart the impacts of solitary confinement and animate absconding patterns has

a growing commercial utility. It is useful, for example, that an archival search can now recover more than links to individual records. Thus, users can also be provided with information about each place the subject of their search was sent to toil. The creation of contextualised finding aids may not be the reason why datasets were originally created, but they might provide a reason for investing in them in the future.

ACKNOWLEDGEMENTS

This research was supported partially by the Australian Government through the Australian Research Council's Linkage Infrastructure, Equipment and Facilities funding scheme (project LE200100074) and Linkage Scheme (project LP180101048).

REFERENCES

- Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal microdata, version 4. *Historical Life Course Studies*, 1(1), 1–29. Retrieved from <http://hdl.handle.net/10622/23526343-2014-0001?locatt=view:master>
- Australian Bureau Statistics. (2014). *Australian Historical Population Statistics*. Retrieved from <https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/3105.0.65.0012014>
- Dunning, T. P., & Maxwell-Stewart, H. (2002). Mutiny at Deloraine: Ganging and convict resistance in 1840s Van Diemen's Land. *Labour History*, 82, 35–47. doi: 10.2307/27516840
- Finnane, M., Kaladelfos, A., Piper, A., Smaal, S., Blewer R., & Durnian, L., et al. (2016). *The Prosecution Project Database*. Retrieved from <https://app.prosecutionproject.griffith.edu.au/>
- Gavrilova, N. S. & Gavrilov L. A. (1999). Data resources for biodemographic studies on familial clustering of human longevity. *Demographic Research*, 1(4), 1–48. doi: 10.4054/DemRes.1999.1.4
- Godfrey, B., Inwood, K., & Maxwell-Stewart, H. (2018). Exploring the life course and intergenerational impact of convict transportation. In V. Eichelsheim & S. van de Weijer (Eds.), *Intergenerational continuity of criminal or antisocial behaviour* (pp. 61–75). London: Routledge.
- Goldin, C., & Katz L. F. (2001). Decreasing (and then increasing) inequality in America: A tale of two half centuries. In F. Welch (Ed.), *The causes and consequences of increasing inequality* (pp. 37–82). Chicago: University of Chicago Press.
- Inwood, K., Kippen, R., Maxwell-Stewart, H., & Steckel, R. (2020). The short and the tall: Comparing stature and socio-economic status for male prison and military populations. *Social Science History*, 44(3), 463–84. doi: 10.1017/ssh.2020.14
- Kellermann, N. P. (2013). Epigenetic transmission of Holocaust trauma: Can nightmares be inherited? *The Israel Journal of Psychiatry and Related Sciences*, 50(1), 33–39.
- Kippen, R. (2011). 'Incorrect, loose and course terms': Classifying nineteenth-century English-language causes of death for modern use. An example using Tasmanian data. *Journal of Population Research*, 28, 267–291. doi: 10.1007/s12546-011-9065-2
- Kippen, R., & McCalman, J. (2015). Mortality under and after sentence of male convicts transported to Van Diemen's Land (Tasmania), 1840–1852. *The History of the Family*, 20(3), 345–365. doi: 10.1080/1081602X.2015.1022198
- Kuzawa, C., & Eisenberg, D. (2014). *The long reach of history: Intergenerational and transgenerational pathways to plasticity in human longevity*. Washington: National Academies Press.
- Maxwell-Stewart, H. (2015). Convict labour extraction and transportation from Britain and Ireland 1615–1870. In C. Vito & A. Lichtenstein (Eds.), *Convict labour: A global regime* (pp. 169–196). Leiden: Brill.
- Maxwell-Stewart, H. (2016). The state, convicts and longitudinal analysis. *Australian Historical Studies*, 47(3), 414–429. doi: 10.1080/1031461X.2016.1203963
- Maxwell-Stewart, H., Inwood, K., & Stankovich, J. (2015). The prison and the colonial family. *The History of the Family*, 20(2), 231–248. doi: 10.1080/1081602X.2015.1006654
- Maxwell-Stewart, H., & Kippen, R. (2015). Sicknes and death on convict voyages to Australia. In P. Baskerville & K. Inwood (Eds.), *Lives in transition: Longitudinal research in historical perspective* (pp. 43–70). Kinston: McGill-Queen's University Press.

- Meredith, D., & Oxley, D. (2005). Contracting convicts: The convict labour market in Van Diemen's Land 1840–1857. *Australian Economic History Review*, 45(1), 45–72. doi: [10.1111/j.1467-8446.2005.00127.x](https://doi.org/10.1111/j.1467-8446.2005.00127.x)
- McCalman, J., & Kippen, R. (2020). The life-course demography of convict transportation to Van Diemen's Land. *The History of the Family*, 25(3), 432–454. doi: [10.1080/1081602X.2019.1691621](https://doi.org/10.1080/1081602X.2019.1691621)
- Nicholas, S. (Ed.) (1988). *Convict workers: Reinterpreting Australia's past*. Cambridge: Cambridge University Press.
- Panza, L., & Williamson, J. G. (2019). Australian squatters, convicts, and capitalists: Dividing-up a fast-growing colonial pie 1821–71. *Economic History Review*, 72(2), 568–594. doi: [10.1111/ehr.12739](https://doi.org/10.1111/ehr.12739)
- Pike, I. (2016). Calibrating the next generation: Mothers, early life experiences, and reproductive development. In L. L. Sievert & D. E. Brown (Eds.), *Biological measures of human experience across the lifespan: Making visible to invisible* (pp. 13–28). Berlin: Springer.
- Riley, J. C. (2001). *Rising life expectancy: A global history*. Cambridge: Cambridge University Press.
- Robson, L. L. (1965). *The convict settlers of Australia: An enquiry into the origin and character of the convicts transported to New South Wales and Van Diemen's Land 1787–1852*. Carlton: Melbourne University Press.
- Rose, M. E., & Cherpitel, C. J. (2011). *Alcohol: Its history, pharmacology and treatment*. Minnesota: Hazelden.
- Sternfeld, J. (2011). Archival theory and digital historiography: Selection, search and metadata as archival processes for assessing historical contextualisation. *The American Archivist*, 74(2), 544–575. doi: [10.17723/aarc.74.2.644851p6gmg432h0](https://doi.org/10.17723/aarc.74.2.644851p6gmg432h0)
- Stoler, A. L. (2009). *Along the archival grain: Epistemic anxieties and colonial common sense*. Princeton: Princeton University Press.
- Tuffin, R., & Gibbs, M. (2019). Repopulating landscapes: Using offence data to recreate landscapes of incarceration and labour at the Port Arthur penal station, 1830–1877. *International Journal of Humanities and Arts Computing*, 13(1–2), 155–181. doi: [10.3366/ijhac.2019.0234](https://doi.org/10.3366/ijhac.2019.0234)
- Tuffin, R., & Gibbs, M. (2020). *Convict landscapes: Locating Australia's convicts, 1788–1868 — Van Diemen's Land*. Retrieved from www.convictlandscapes.com.au/VDL
- Tuffin, R., Gibbs, M., Roberts, D., Maxwell-Stewart, H., Roe, D., Steele, J., & Hood, S., (2019). Convict labour landscapes, Port Arthur 1830–1877. doi: [10.25952/5de58b5512209](https://doi.org/10.25952/5de58b5512209)
- Tuffin, R., Maxwell-Stewart, H., & Quinlan, M. (2020). Reintegrating historical records through digital data linking: Convicts prosecuted for collective action in Van Diemen's Land. *Journal of Australian Colonial History*, 22, 49–84.
- Tuffin, R., Roe, D., Gibbs, M., Clark, D., & Clark, M. (2020). Landscapes of production and punishment: LiDAR and the process of feature identification and analysis at a Tasmanian convict station. *Australian Archaeology*, 86(1), 37–56. doi: [10.1080/03122417.2020.1749406](https://doi.org/10.1080/03122417.2020.1749406)
- United Nations Office of Drugs and Crime. (2015). *International classification of crime for statistical purposes (ICCS), version 1.0*. Retrieved from <https://www.unodc.org/unodc/en/data-and-analysis/statistics/iccs.html>
- van Leeuwen, M. H. D., & Maas, I. (2011). *Hisclass: A Historical International Social Class Scheme*. Leuven: Leuven University Press.
- van Leeuwen, M. H. D., Maas, I., & Miles, A. (2002). *HISCO: Historical International Standard Classification of Occupations*. Leuven: Leuven University Press.

APPENDIX

Dataset	Archival Reference	Date Range	Key variables	Number of records
3.1 Convict				
Grange Gorman Female Penitentiary (Dublin, Ireland)	National Archives of Ireland: MFGS 51/028	1840–1852	Prisoner number Name Age Hair colour Eye colour Height (inches) Literacy Marital status Religion Trade Previous convictions Transportable offence Trial date Court Sentence Date received Date discharged How disposed of	3,594
Pentonville register (London, UK)	National Archives (UK) PCOM 2/61	1842–1847	Prisoner number Name Age Weight (pounds) Date received Place received from Place of birth Father's name Father's occupation Father's address Prisoner's marital status Number of children Trial date Court Sentence Conduct in prison Date discharged How disposed of	1,821
Millbank register (London, UK)	National Archives (UK) PCOM 2/21	1837–1845	Prisoner number Cell number Name Hair colour Eye colour Height (inches) Age Number of children Religion Date of conviction Place of conviction Crime Sentence Date received Character Date discharged How disposed of	5,688

Hulk Registers (UK)	National Archives (UK) HO/9 series	1833–1844	Hulk Record number Committing gaol or hulk Date received Name Age Crime Conviction place Date of conviction Sentence Literacy Marital status Trade Gaol report When disposed How disposed of	36,643
Surgeon's sick list Voyage to Australia	National Archives (UK) Adm/101 series	1817–1853	Ship Date of sailing Date patient entered on sick list Name Age Date discharged How disposed of	24,340
Description lists for male and female convicts	Tasmanian Archives Con 18, 19 and 23 series	1816–1853	Police number Name Ship of arrival Date of arrival Trade Height (inches) Age (on arrival) Hair colour Eye colour Remarks (including tattoos and injuries)	77,330
Appropriation registers for male and female convicts	Tasmanian Archive Con 14 and 15 series	1825–1844	Police number Name Ship Date of arrival Age on arrival Trade Person on government department convict ordered to work for	11,573
Conduct records for male and female convicts	Tasmanian Archive Con 31, 32, 33, 40 and 41 series	1816–1853	Police number Name Ship Date of arrival Where sentenced Date sentenced Sentence Transportation offence Gaol report Hulk report Marital status Convict's version of transportable offence and confessed prior convictions Details of any family (children, brothers and sisters, etc.) Surgeon's report (voyage to Australia) Religion Literacy	49,248

Colonial Prosecutions for male and female convicts	Tasmanian Archive Con 31, 32, 33, 40 and 41 series	1816–1902	Police number Name Ship Date of trial for subsequent colonial prosecutions Location where convict deployed at time of offence Description of charge Sentence Magistrate's name	117,943
Permission to marry register	Tasmanian Archive Con 45 & 52 series	1829–1860	Name of individual lodging application Ship of arrival Name of intended spouse Ship of arrival of spouse Date application was lodged Application outcome	29,313
Gazette Notices	Hobart Town Gazette	1817–1860	Police number Name Ship Date convict absconded Date recaptured Date transferred between employment locations Date of appointment to police Date appointment revoked Date probation completed Date awarded ticket of leave, certificate of freedom or pardon Date ticket of leave revoked	286,679
Pass Holder contracts	Hobart Town Gazette	1844	Police number Name Ship of arrival Name of employer Place of employment Length of contract	7,512

3.2 Colonial Courts & Criminal Justice

Coal Mines	Tasmanian Archive AF584/1/1	1836–1841	Police number Name of offender Ship of arrival Date of trial Offence Sentence magistrate	1,645
Hobart Petty Session (cases brought against women)	Tasmanian Archive LC251 series	1846–1854	Police number Name of offender Ship of arrival Literacy Name of employer Place of employment Offence Plea Sentence Witness statement Name of magistrate	5,086

Descriptions of convicted & discharged prisoners	Tasmania Police Gazette	1865–1924	Location where prisoner was discharged Name of prisoner Sex Alias Ship to colony Where tried Date tried Offence Sentence Native place Age Height (inches) Hair colour Eye colour Remarks including civil status (free by servitude, came free, colonially born) and distinguishing marks	50,387
--------------------------------------------------	-------------------------	-----------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------

3.3 Census & Musters

Convict musters	National Archive (UK) HO 47, 48, 49, 50	1830, 1832, 1833, 1835	Police number Name Ship Place convict located	59,948
Convicts resident in Richmond Police District	POL 584 series	1831–1834	Police number Name Ship Place convict located General conduct	2,572
Census manuscript returns	Tasmanian Archive Cen 1 series	1837–1857	Parish Name of household head Address Name of proprietor Property description Number of residents Age and sex of residents Status of residents (convict or free) Religion Occupation	14,870
Hobart street & trade directories	Various almanacks	1825–1854	Name Occupation or business address	25,449

3.4 Free Departures & Arrivals

Free arrivals	Tasmanian Archive arrival index	1817–1858	Name Title Ship of arrival Port of departure Date of arrival	42,232
Departures	Tasmanian Archive departures index	1817–1858	Name Title Ship of departure Port of departure Date of departure Where bound Ship to colony Status (e.g. former convict)	114,452

Assisted migrant arrivals	Tasmanian Archive CB7-1-13-20	1852–1858	Ship Port of arrival Date of arrival Name Age Sex Marital status Religion Native place Literacy Occupation Employer's name Wages Period of hire	10,631
---------------------------	----------------------------------	-----------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------

3.5 Births, Deaths & Marriages

Burial register	Tasmanian Archive RGD 34 series	1803–1838	Name of deceased Sex Place of residence When died When buried Age Ship of arrival Cause Occupation By whom ceremony performed	8,828
Baptism register	Tasmanian Archive RGD 32 series	1803–1838	Name of child Name of father Name of mother Date of baptism Place of baptism Denomination (Anglican, etc.) By whom baptised	19,723
Death register	Tasmanian Archive RGD 35 series	1838–1928	District of registration Date of death Name Sex Age at death Occupation of deceased Place of birth Cause of death Name of informant Residence of informant Date of registration Name of registrar	155,000
Birth register	Tasmanian Archive RGD 33 series	1838–1899	District of registration Rank or occupation of father Date of birth Name of child Sex of child Name of mother Maiden name of mother Name of father Name of informant Description of informant Residence of informant Date of registration Name of registrar	195,000

Marriage register	Tasmanian Archive RGD 36 & 37 series	1838–1899	District of registration Date of marriage Place of marriage Name of groom Name of bride Age of groom Age of bride Name of clergyman Date of registration Name of church or chapel Locality of church or chapel Rites used Signature or mark of groom Signature or mark of bride Name of witness 1 Signature or mark of witness 1 Name of witness 2 Signature or mark of witness 2 Marriage by licence or certificate	51,000
-------------------	-----------------------------------------	-----------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------

3.6 Property & Financial Records

Money & property deposited by convicts	Tasmanian Archiv Con 73, 122 & 147; Derwent Bank boxes: 15b7; 16b2; 18b9; 18b11; 19b6; and TAHO, GO33: 1/14. State Library NSW, Tas Papers 21–26	1829–1860	Police number Name Ship Amount deposited Description of other property held in trust by the state Date deposited Amount withdrawn Date withdrawn	24,554
Hobart Savings Bank	Tasmanian Archives NS1167	1845–1863	Year Name Sex Whether signed or marked with a X Employment Address Age Height (inches) Physical description Ship to colony	12,240
Valuation rolls	Hobart Town Gazette	1853–1883	Location of property Description of property Area (acres, perches, rods) Valuation Occupier Place of residence of occupier Owner Place of residence of owner	153,291

3.7 Hospital & Pauper Records

Hobart hospital deaths	Tasmanian Archive HSD 145 series	1864–1884	Name of deceased Ship of arrival Place of birth Age Religion Occupation Date of death Cause of death Date of burial	2,650
------------------------	-------------------------------------	-----------	---------------------------------------------------------------------------------------------------------------------------------------------	-------

New Norfolk asylum admissions	Tasmanian Archive Con 127 series, HSD 246 series, HSD 247 series	1830–1899	Name Ship of arrival Date of admission Place admitted from Age on admission Diagnosis Date of discharge	5,720
Pauper admissions	Tasmanian Archive NS 1172, HSD 274 series, POL 709 series, SRCT series	1858–1952	Name Ship of arrival Name of institution Age on admission Date of admission Date of discharge	35,161

3.8 Military Records

WWI attestation papers	Australian National Archive, MT1486 series	1914–1918	Name Next of kin Place of birth Place of enlistment Date of enlistment Age on enlistment Trade Prior rejection Marital status Number of children Height (inches) Weight (pounds) Expanded and unexpanded chest (inches) Eye colour Literacy Religion Details of discharge Date of discharge	15,234
------------------------	--------------------------------------------	-----------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------

3.10 Meteorological Data

	Hobart Town Gazette, Colonial Times, Courier, Hobart Town Daily Mercury, Mercury	1825–1827, 1838–1879	Date Min temperature Max temperature Min barometer Max barometer Wind direction Weather observation, eg. cloudy, drizzle, rain	8,800
--	----------------------------------------------------------------------------------	-------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------	-------

Construction of the Finnish Army in World War II Database

Ilari Taskinen

Tampere University & Visiting researcher Radboud University Nijmegen

ABSTRACT

This article introduces the Finnish Army in World War II Database (FA2W) currently under construction that is being built to study the effects of World War II on Finnish society. The database is a stratified sample of 4,253 representative of the men who served in the Finnish Army in World War II. The data have been gathered from the military service record collection of the Finnish Army, which holds files on practically all draft-age Finnish men of the birth cohort 1903–1926 and around 70% of the birth cohorts 1897–1902. The amount of data is extensive, containing over 60 different variables. The main part of the database consists of men's military careers, comprising longitudinal data on their positions in society and in the army (e.g., civilian/conscript/frontline service), military unit, military branch, task, rank, and service class. Other information includes socio-economic information from the draft and wartime and war experiences, such as wounds, illnesses, medical treatments, death, and honors. In the future the database will be expanded with men's postwar life trajectories to study the long-term effects of the war.

Keywords: Historical database, Demography, Soldiers, Mortality, Social class, World War II, Finland

DOI article: <https://doi.org/10.51964/hlcs13565>

© 2023, Taskinen

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>.

1 INTRODUCTION

Experiences of wartime violence have severe demographic consequences. Most obviously, many people die prematurely, and the effect of these deaths on society has been studied by social historians for decades (Urlanis, 1971; Winter, 1986). More recently, it is better understood that wars severely impact the life course of survivors, as the long-term effects of violence, such as psychological trauma, have emerged in both academic and public discussions (Kivimäki, 2013). The study of these consequences has largely been conducted using contemporary sources in fields like health sciences and sociology (MacLean & Elder, 2007; Modell & Haggerty, 1991). However, as the immensely successful Union Army project on American Civil War soldier data (Fogel et al., 2000) and recent projects in Italy (Fornasin, Breschi, & Manfredini, 2019) and Australia (McCalman, Kippen, McMeeken, Hopper, & Reade, 2019) have shown, demographic historians with their vast historical datasets also have much to offer for the analysis of the human consequences of wars.

Human effects of war were in the focus of the project "Large Databases in Studying the History of War Experiences" (STASKO) led by Ville Kivimäki and conducted in Tampere University, Finland in 2017–2020. The project used digital technologies to create new large-scale datasets for the historical research of World War II Finland. One of its inspirations was the data on the Finnish soldiers who died in that war. The documentation on these approximately 94,000 men is excellent due to practices initiated in wartime. It was exceptional for a nation at war that the Finns did not bury their dead on the battlefield, but brought them back to their hometowns, where they are now at rest in the so-called "hero graves" in prominent locations outside every Finnish church (Kemppainen, 2006). Documentation on these Finnish casualties of war is currently stored in a public database containing socio-economic, military, and cause of death information on each of the fallen (see <https://www.sotasampo.fi/en/casualties/>). Construction of this highly comprehensive database was begun in the 1980s by private individuals, and it has been managed and updated since the 1990s by the National Archives of Finland. Since 2000, data on fallen soldiers has been available on their internet service, and the raw data is currently available to all interested parties (Ahoranta & Ortamo-Närvä, 2012; Karjalainen, 2014).

The database of the fallen enables statistical analyses of people who died in the war. However, studies of differential mortality (e.g., by social class) are limited by the absence of data on soldiers who survived the war and on men who saw no active service. In the STASKO project we investigated the possibility of taking a representative sample of all Finnish men of the age cohorts that participated in the war. After investigating the military service records of the Finnish Army, we found that this was indeed possible, and this article describes the Finnish Army in World War II Database (FA2W), which is currently under construction. The database consists of a representative sample of 4,253 of Finnish men in the birth cohorts who fought in World War II. The database is based on the military records of the Finnish Army, and contains extensive information on soldiers' social backgrounds, health, military service, and war experiences. In the future the database will be expanded with postwar data. The database is important for understanding the effects of World War II on Finnish society, and it offers excellent opportunities for studying the consequences of war on the life course. Finland may have the most comprehensive and detailed archives on World War II soldiers' wartime service and post war experiences.

2 HISTORICAL BACKGROUND

Before turning to the database, I will briefly review the history of Finland in World War II. This is beneficial for understanding the premises of the database and the opportunities it affords for the demographic study of war.

On the eve of World War II Finland was a small nation of four million people at the top of northern Europe. The nation had been part of the Kingdom of Sweden from the 14th century until 1808 and a grand duchy of the Russian Empire in the period 1809–1917 before gaining independence in 1917. In World War II, the young nation fought three wars. The first of these was the Winter War, which began on November 30, 1939, when the Soviet Union launched an attack to invade her small neighbor which had fallen under her sphere of influence in the secret protocol of Molotov-Ribbentrop Pact signed between the Soviet Union and Nazi Germany. Against all odds, Finland was successful in defending herself and in March 1940 signed a peace treaty after three and a half months of fighting. The nation ceded 10% of its territory in the eastern parts of the country but retained its independence.

Hostilities were resumed when Finland joined Germany to attack the Soviet Union in June 1941 in Operation Barbarossa. This war, known as the Continuation War in Finnish historiography, lasted until September 1944 and saw a successful Finnish offensive in 1941, two and a half years of relatively calm stationary warfare from winter 1942 to summer 1944 and an intense Soviet offensive in 1944, which Finland again warded off. An armistice was signed in September 1944, which led to the third war, the Lapland War, fought from fall 1944 until spring 1945 in northern Finland between the Finns and their former allies, the Germans (Kinnunen & Kivimäki, 2012).

Those familiar with the history of World War II will notice the unique situation of Finland: unlike all other small nations of Eastern Europe, Finland was never occupied. This exceptional and fortunate fate is the basic component of Finnish World War II demographic history, and its consequences are apparent in the nature of the casualties: between 1939 and 1945 Finland lost around 94,000 soldiers but only 2,000 civilians due to the war (Kivimäki, 2019). This ratio is in stark contrast with casualty figures in other European countries, where, without many exceptions, tens of thousands to millions of civilians perished under total warfare and systematic murder (Bessel, 2015). In Finland, however, the war mostly spared civilians. The Soviet Union bombed Finnish cities, but the destruction of these raids was relatively limited and the civilians could continue their lives mostly in safe circumstances, albeit under great stress and material shortages.

This leads us to Finnish men, the subjects of the database described in this article. While the war spared Finnish women and civilians, it was experienced firsthand by a great portion of Finnish men. In order to survive, the small nation was compelled to mobilize the maximum number of men for military tasks (Kurenmaa & Lentilä, 2005). The exact number of mobilized men is one of our research issues, but according to preliminary results, it seems that as many as 750,000–800,000 Finnish men served between 1939 and 1945. This is a considerable number for a nation of four million and means that over 80% of Finnish men of the birth cohorts 1897–1926, the primary age groups from which Finland mobilized men, took part in the war. This extensive participation is a fundamental aspect of the Finnish demographic history of World War II. While in many other countries combat touched only certain social groups, in Finland the war was inherently an experience of the whole nation. The Finnish Army in World War II database reveals how a nation and all its social strata experienced and suffered from the consequences of war.

3 DATA: MILITARY SERVICE RECORDS

Studying the life courses of soldiers and other people in or after a war involves numerous challenges. Often, an obstacle is the lack of nationwide registers, and even in cases where such existed, they may have survived only partially due to inadequate archival practices or destruction caused by outside forces, a danger severely heightened in wartime. In this framework, the opportunities to conduct demographic research in Finland are exceptional thanks to the military service records of the Finnish Army. These records are the basis of the FA2W database, which I will introduce in this section. In the future the database will include data on soldiers' post-war lives, but the current first stage of the database construction focuses on soldiers' pre-war and wartime lives.

Both the survival and richness of the military records is a consequence of Finland's wartime fate. Unlike many other countries, where military archives were lost and destroyed in combat and invasions, in Finland these losses were few because the country was never occupied and not even the Russian bombing of the Finnish home front did much damage. However, Finland needed to make the greatest possible use of her small population to survive, in practice calling to arms every man it could spare. This would not have been possible without meticulous recordkeeping.

The Finnish state adopted universal male conscription after gaining independence in 1917. Between 1918 and 1939, the Finnish Army trained for reserve 509,000 men, which was around 70% of the age cohorts that were called up in drafts (Kronlund, 1988). However, a greater number of men ultimately came into contact with the army because, in desperate need of manpower, the army called up previously exempted men to redrafts during the war. Exemptions from military service were granted only to the very weakest and sickest (Nurminen, 2008).

Information on Finnish men was recorded in various documents during their basic training and wartime service. These records are now organized into several different collections, of which two are used in the construction of the FA2W database. First are the draft records, *kutsuntaluettelot*, compiled for the annual drafts of the army. These drafts were based on information provided by Finnish population register keepers, most notably the Evangelical Lutheran Church of Finland, to which over 95% of the Finnish people belonged, and by other churches and civilian register keepers. Each year before the drafts these authorities provided the regional military authorities with information including the names, places of residence, occupations, education, and marital status of all men of draft age. They compiled this information into draft district specific lists, which were used to call up the men and were updated with additional information such as men's height and weight and the results of draft.¹ These records are currently archived in the National Archives of Finland and, as I will later discuss, they are used as a supplementary source when the corresponding data is missing from the main sources of the FA2W database.

The second collection, and the one that is the base of the FA2W database, is the military service record collection of the Finnish Army. This collection of personnel files contains various documents, although it is mostly known for its main document, the military service record *kantakortti* (see Figure 1). This is a two- to four-page record, depending on its form, which was updated in 1930 and 1945. Every Finnish man who was called up in a draft got such a form which was updated throughout his military service. This record contains extensive information including socio-economic details (e.g., occupation, education, marital status), military training (e.g., military branch, education, evaluation), military service history (units, ranks, tasks, service class) and war experiences (e.g., wounds, illnesses, battles, honors). For officers an additional form contained some further information, such as evaluations of an individual's suitability for different tasks and the occupation of the father.

In addition to the military service record, the personnel files include a wide variety of other documents. A medical inspection record was in principle created for every man in the draft and subsequently updated during his service. This document includes information on medical examinations (e.g., previous illnesses, height, weight), customary procedures (e.g., vaccinations), illnesses, and treatments. Medical information can also be found in individual documents written in military and field hospitals. These hospitals produced reports on the diagnoses and treatments of men which were attached to their personnel files. The files also include a wide variety of other documents gathered by the army, such as disciplinary records and information on their deaths.

The amount of data available in the military service files is vast. On average, there are 14 pages of documentation per man in the sample (see next section), which is an underestimation for men in active service, because those who were exempted may have only two pages. Men with long and eventful military careers may have over 50 pages of documentation. This plethora of data is not merely due to the medical appendix but also to different versions of the military service records. The original pre-war recording principle was to have two copies per man, one of which was kept at the military district headquarters while the other followed a man to his various military units, where it was updated with his service details. When a man completed his military training or returned home during the war, the record was returned to district headquarters, where the more recent information was copied into the main document (Ylönen-Peltonen, 2021). This system seems to have been too difficult to implement in the chaotic circumstances of the war, when men were rapidly transferred between units. When a man joined a new unit during the war, a new service record was often begun. This is fortunate for us, because military service files often include records from men's prewar conscription service as well as from the war, enabling us to gather data on socio-economic status from both periods.

In the Finnish Army, the military districts were responsible for the mobilization of the detachments of the army in the event of war. Military districts kept records of the men on active service and the reservists through the military personnel files. If a man transferred between military districts, his file followed, and this continued for the duration of his status as a reservist. Unless reserve status was discontinued due to poor health, men could stay in the army reserves until age 60, and some information in these files was updated during this time. Since World War II, Finland has not been at war and most men of the war generation did not serve in the army again, except for some refresher training, so postwar documentation is sparse. However, changes in places of residence, information on deaths and social security numbers, introduced into Finland in late 1960s, were frequently updated in the postwar decades.

1 The National Archives of Finland, PLM-33/Ee:3-Ee:5, The draft district regulations and the draft district instruction from the Ministry of Defence, 4 December 1929.

For the youngest men of the war generation, born in the 1920s, the service records may contain entries until the 1980s. After a man was removed from the army's reserves, the service files were first sent to the Central Medical Archive of the Finnish Army, and they were moved in the 1990s to the Military Archive of Finland, which is today part of the National Archives of Finland (Ylönen-Peltonen, 2021).

The meticulousness with which the military service records were stored and updated reveals the importance that Finland, a small nation next to a great superpower, attached to securing her limited population for any national defense effort. The continued importance of this is reflected in the fact that Finland is one of only two nations in Europe to retain universal conscription. At the same time, these archival practices indicate why the military service record collection is not only rich in content but also comprehensive. According to research conducted for the building of FA2W database, the National Archives of Finland stores military service files on nearly every Finnish man born between 1903 and 1926 who survived to draft age. For men born 1897–1902, around 70% were included, omitting 30% who were exempted from army service. Information on this exempted group can be gathered from the draft list to obtain a highly representative sample of all Finnish men of the age cohorts who fought in the war.

Figure 1 The four-paged model 1945 military service record (kantakortti).

Explanation: The name, date of birth, and social security number of the person are not exposed. The first page contains personal, socio-economic, and conscription service information, the second page career in military service, the third lists promotions, honors, service classes, wounds, illnesses and battles, and the last page contains information on e.g., offences.

4 SAMPLE IN TWO STEPS

The FA2W database is based on a stratified random sample of 4,253 men taken from the military service record collection and the draft lists of the Finnish Army. At the beginning of our project, we lacked a comprehensive overview of the content of the military service record collection, thus, we initially took a less accurate, systematic random sample. After gaining detailed knowledge of the military service record, we could reshape the sample into a more accurate stratified sample.

According to the army guidelines from the early 1930s, a military service record was to be written up for every Finnish man called up in a draft. This meant that a record should have been made for every Finnish man who was alive and of the relevant age because Finland practiced strict universal conscription (Ahlbäck, 2014). When we began to plan a sample from the military service record collections, it was not known how well this principle had been followed and to what extent the records were extant. The archive knew that the collection was very comprehensive and holds altogether 4.5 shelf kilometers of documents. It is the most used collection of the National Archives due to its wide usage in genealogical research (Ylönen-Peltonen, 2021). However, the archive also warned us that some men were missing from the collection. These were presumably most often men of the older age cohorts who had been exempted in drafts or during their conscription service.

At this point there was no way to investigate these limitations in detail. It would have been all but impossible from the kilometer-long rows of file stacks, and it was also difficult from the otherwise very helpful indexes of the collection. The indexes were organized in two parts: a physical one (see Figure 2 for an example) and an electronic one. The electronic file contained only the names of 80% of the men born during 1924–1926. The physical card index contained men born 1897–1923 (and a small number of men from earlier age groups) and remaining 20% of the age cohorts 1924–1926. These cards were kept in 779 boxes each holding around 1,250 cards. The index was in principle organized by birth year and in alphabetical order according to the family name within years (see Table 1). However, the age cohorts 1900–1909, 1910–1918, and all men born before 1898 were organized into separate series, and around one-third of men killed in the war had their own series.

Figure 2 *Military record card index box and index card*

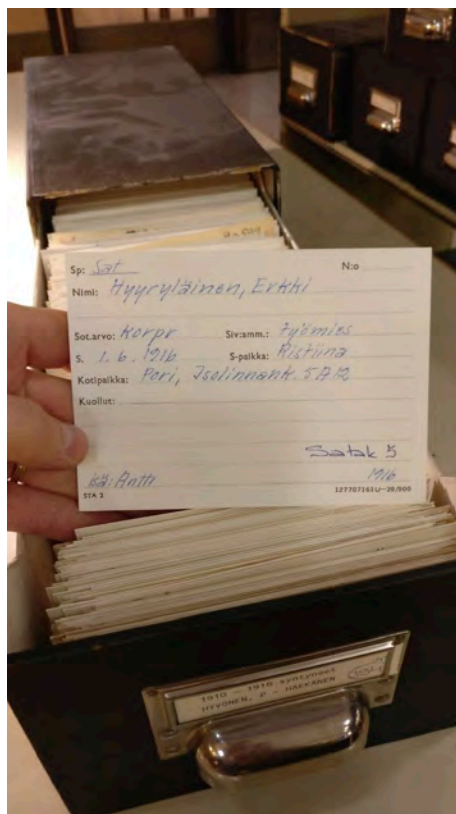


Table 1 *Organization of the indexes of the military service record collection*

Series	Card index		Electronic database	
	Boxes	Age cohort	Men	
-1897	59			
1898	19			
1899	19			
1900–1909	248			
1910–1918	262			
1919	26			
1920	27			
1921	27			
1922	27			
1923	30			
1924	6	1924		26.996
1925	5	1925		28.612
1926	4	1926		29.319
1st series of fallen	20			
2nd series of fallen	66*			

* Cards of the 1st series of the fallen, holding around one third of the soldiers killed in the war, have been collected from the age cohort series at some point after the original assembly of the card index. After we gathered the initial sample, the archive collected the remaining casualties of the war to a second series of fallen. This series is packed into different, smaller boxes than the rest of the card index.

Due to these exceptions, it was not possible to make a stratified sample, e.g., to ensure enough cases from each age cohort. We therefore opted for systematic random sampling. This would provide us with a representative sample of the Finnish men who served in the Finnish Army in World War II, under the assumption that soldiers' records would be better preserved than those exempted from the draft. The sample was taken by 1) picking randomly five cards from each of the 779 boxes of the physical card index and 2) picking 336 men from the electronic database, based on the sampling interval calculated by counting the number of cards in the physical card index boxes. After excluding unnecessary reference and duplicate cards picked from index cards, the result was a sample of 4,045 men.

After this initial sampling we started to enter the data, and after completing a quarter, we could analyze the representativeness of the sample. Here we made two important observations: one was that there was a major increase in the number of men in the sample between the age cohorts 1902 and 1903, which suggested that there had been a change in the principle of the creation or preservation of records between these years. The other observation posed a problem for our sample. When comparing the distribution of soldiers killed in the war in our sample with those from the database of the fallen compiled by the National Archives of Finland, we found that there were too many of the fallen in the older age cohorts of the sample and too few in the younger age cohorts.

This observation led us to do further research on the indexes of the collection, which at this point was much easier thanks to the work accomplished by the National Archives of Finland. After our initial sampling, the National Archives of Finland had drawn the index cards of the fallen men still in the main index into their own boxes and added the information on these cards to an electronic database. This new organization of the information revealed, first, that the men included in the database of the fallen were well represented in the military service record collection: around 99% of them had a card in the index and around 97% of their corresponding documents could be found, while the remaining 2% seem to have been lost during archiving. Second, it was now possible to estimate the number of survivors and the total count of men in the index card series. For this, we manually measured the

length of the card rows in 84 card boxes and counted the number of cards in 22 boxes to estimate the number of cards in the index series (see Figure 3).

This new measurement had two main results. First, it revealed the reasons why our initial sample had been distorted. We had erroneously assumed that the card boxes held a consistent number of cards across cohorts, but the measurements revealed discrepancies of 10 to 20% in their card counts. The main reason for the discrepancies was that the series containing one third of the fallen had been collected unevenly, leaving different counts of cards in the age cohort series. Moreover, part of the age cohort series seems to have been condensed into a smaller number of boxes at some point, and these boxes were very tightly packed with relatively large numbers of cards. These findings and the observation that the index cards of the fallen were of a slightly different size than the other cards explained the distortions in our sample.

Second, our measurements of the index cards gave us data about the numbers of men in the military service record collection. As noted, this could be done precisely for the men who died in the war, whose records were confirmed to have been almost entirely preserved. The collection is also very comprehensive for other men. According to our estimates, most age cohorts actually had more cards in the index than the number of Finnish men alive and of draft age in these cohorts (see Table 2). The index included extra cards, such as men who had changed their names. In our initial sample, 4.4% of the cards were of this kind. But even taking these cards into consideration, the collection is very comprehensive.

Figure 3 *Measurement of the length of index card stack*



Measurement of the card index disclosed that the military service record collection was highly representative for the majority of the age cohorts serving in the war. Except for the series of the 1919 age cohort, which curiously contains 12% more index cards than the draft-age population of the cohort, the cohorts 1910–1926 differed only marginally from the expected 100%. Taking account of the imprecision of our measurements, it is safe to say that the military service records were written and archived very meticulously. The small number of missing records should not cause large distortions to the collection.

However, the series 1900–1909 show a notable deficit with 12% of the men missing. This series also marked a significant turning point in the collection as the numbers of men in our sample increased markedly between the age cohorts 1902 and 1903. It seems that military service records were written and preserved on nearly all Finnish men from the age cohort 1903 onwards, although some men who were exempted from drafts might still be missing. Among the older age cohorts, the military service records are missing from around 30% of the men. Our analysis of the initial sample indicates that these were men who had been exempted from the draft or from conscription service, but this was not a strict rule as some of the exempted men did end up in our sample.

Table 2 *Percentages of men in the military service record collection as part of the total number of conscripts per birth year*

–1897	70,115*
1898	73%
1899	72%
1900–1909	88%
1910–1918	103%
1919	112%
1920	100%
1921	101%
1922	96%
1923	99%
1924	103%
1925	102%
1926	101%

Source: *Finnish Population Statistics and sample results.*

* *It was not possible to make the calculation of the –1897 series as it contains men from unknown age cohorts. The estimated number of men in this series is 70,115.*

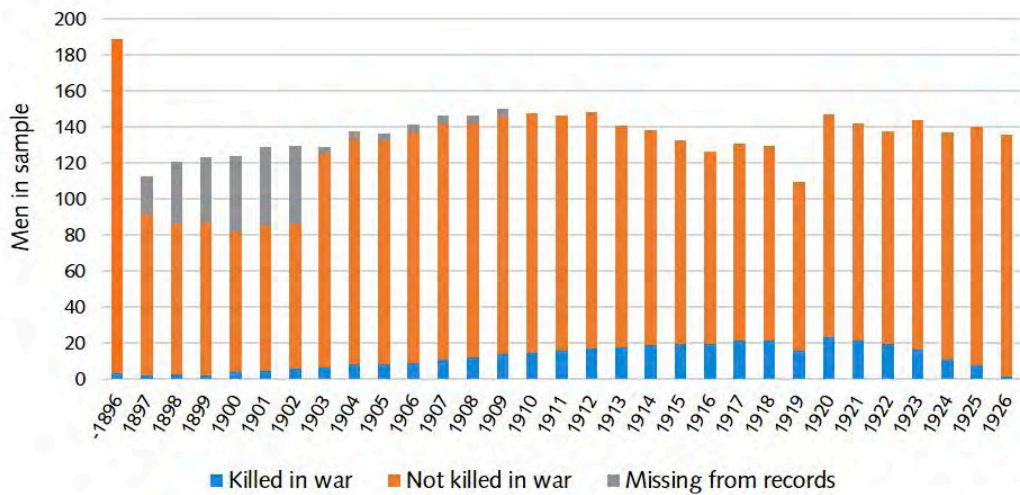
With this research we could now divide the men in the military service record collection into those who died and those who survived the war, and we also knew how many and which types of men were missing from the collection. This new information enabled us to design a better, stratified sample, which would not be based on the record collection like the initial sample, but on population statistics. The new sample was stratified on thirty age cohorts for birth years 1897 to 1926 using the size of each cohort at end of the year preceding the draft.² Men in older birth cohorts with a file in the military service record collection were also added to the sample. The size of this sample population is 1,063,141.

Men were further divided into two primary strata within the birth cohorts: those who were killed in the war and those who survived. A third stratum of men who were missing from the military service record collection was added to the birth cohorts 1897–1909. These men were exempt from military service and did not get a military service record. However, they are recorded in draft records (*Kutsuntaluettelot*), introduced in the previous section, from which we will sample these men in the near future. There were 75 strata in all as shown in Figure 4. The sizes of these strata were calculated with information from the database with war deaths compiled by the National Archives and our measurements on the number of men in the series of the military service record collection. The sizes of the strata were calculated using a sampling interval of 1 in 250, which yielded a sample of 4,253 men.

The men selected for our original systematic random sample served as the basis of the stratified sample. When the original sample for a certain stratum did not correspond with the new stratified sample, we removed or added deceased and surviving men to and from the birth cohorts of the sample. Men removed from the sample were randomly selected from the database. Men were added by random selection from the database of fallen soldiers and the card index boxes. Samples for strata with men missing from the service record collection (1897–1909) will be taken from draft records after the new sample is examined to determine why men exempted from the army did not get a military service record.

2 The Finnish Statistical Office reported age cohort specific population statistics that were based on censuses in major cities and towns and civil register-keepers reports in rural Finland during this period once in 10 years (1920, 1930 and 1940). I have calculated cohort sizes between these years by subtracting from these numbers the age cohort specific yearly mortality figures that were reported every year.

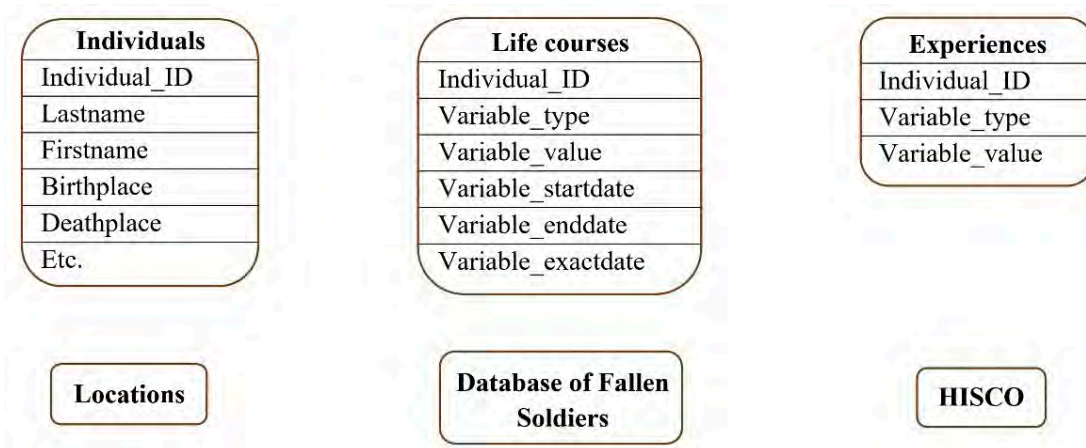
Figure 4 Strata of FA2W sample



5 DESIGN OF THE DATABASE AND VARIABLES

During the first stage of the FA2W database construction, we are compiling all core information available in the military service records. The database is broad, consisting of over 60 variables. These include traditional personal and socio-economic information, but its center is in the detailed portrayal of men's military careers and experiences. These variables often consist of standardized values originating from the military classifications of the Finnish Army, but we have also created some ourselves considering our research aims. Ultimately, the data will be stored in a relational SQL database. It is planned that the database would comprise three main tables Individuals, Life courses and Experiences. Furthermore, we will link these tables to three other datasets (see Figure 5), for coding and standardization (Locations for locations, HISCO for occupational titles) and complementary information (Fallen soldiers). The tables Life Courses and Experiences are designed according the principles of the the Entity Attribute Value model (EAV), in which each record entails only one attribute (Stead, Hammond, & Straube, 1982).

Figure 5 FA2W database design



In the following three tables we explain the variables that are included in the three main tables. The table Individuals contains data without an indication of time. The table Life courses includes dynamic data or static data that have a date like birth_date. The table Experiences contains more complex data and variables summarizing values from the first two tables.

Table 3 *Personal and socio-economic data in the FA2W database*

Static data (not dated)	Events with dates	Attributes at the time of draft/conscription	Attributes during the war years 1939–1945
Last Name	Date of Birth	Place of Residence	Place of Residence
First Name	Date of Draft	Civil Status	Civil Status
Place of Birth	Date of Follow-up Examination	Number of Children	Number of Children
Place of Death	Date of Death	Occupation	Occupation
Legitimacy	Illness	Education	Education
Father's Name	Wound	Height	Height
Mother's Name	Offence	Weight	Weight
Mother Tongue	Sentence	Name of Next of Kin	Name of Next of Kin
Religion	Honor	Relationship of Next of Kin	Relationship of Next of Kin
Nationality		Place of Residence of Next of Kin	Place of Residence of Next of Kin
Foreign Language Skills			
Color of Eyes			
Color of Hair			
Social Security Number			

The table "Individuals" includes personal and socio-economic data. This is a long-format table without time stamps and the variables are static (time-constant), see Table 3, first column. The other columns in Table 3 show the variables that are stored in the table "Life Courses" and which are dynamic in time. The ones in the second column are day-specific events. This includes vital events of birth and death and various events of military service starting from draft and ending in wartime experiences, such as wounds, illnesses and honors. In addition to date, we gather values (e.g. type of illness) for these military experiences. In the first phase of the database building, we gather only serious wartime wounds and illnesses requiring hospital treatment. The durations of treatment periods can be obtained from the variable "Position" (see Table 4).

The variables in the third and fourth column are dynamic in a very limited sense, since they are collected from the military records which make only a distinction between the period of conscription service in peace time and serving during war years. So, as this information was not dated in military service records and often updated during service, the starting date and ending date of each type of period is the best accuracy with which the information can be recorded. Subsequently, the time range of men's conscription and wartime service can be calculated from the table "Life Courses". The variable 'Place of Residence' was systematically updated after the war and then the record includes exact dates.

Table 4 *Military career data in the FA2W database*

Periodic Variables	
Variable	Value
Position	Civilian/Reservist/Conscript/Army Personnel/Field Army/Home Front Troops/Backup Reservist in Service/Treatment in Military Hospital/Furlough/Secondment/Prisoner/Prisoner of War/Deserter/Backup Reservist (Classes I, II & III)
Military Unit	Description
Military Branch	Infantry/Field Artillery/Signals/Coastal Artillery/Sapper/Anti-Aircraft/Military Engineer/Home Front Troops/Armored Force/Logistics/Navy/Air Force
Task	Description
Military Rank	Finnish Army Ranks
Service Class	A I, A II, B I, B II, C, D, E

The core part of the table "Life courses" comprises six variables (see Table 4) that form periods in military careers, and for which the start and end dates are exactly dated. "Position" is the main variable of the group consisting of 16 values designed by us to describe a man's position in society and in the army. This variable can be used to distinguish between civilian and military roles. The most important distinction among men serving in the army is between combat roles in the field army and auxiliary roles on the home front. There are also more specific subgroups, such as those in military hospitals and deserters, that offer interesting options for analysis. The "Position" variable builds an uninterrupted life course for men in the sample. Each day of their civilian lives and military careers, they belong to a single category of the "Position" variable.

These military roles can be further specified with the variables "military unit", "military branch", and "task". The first two are crucial because they divide men into units, like infantry, air force, and logistics, that operated in different wartime environments and faced different dangers. "Task," which includes roles like "rifleman", "squad leader", "clerk", and "driver," identifies further differences within military branches. This is particularly important for the infantry, because it distinguishes the men who really did the fighting in the trenches, like riflemen, from the personnel who worked behind the lines, like clerks. "Military rank," which offers a social and hierarchical perspective on work and roles in the army, and "service class," which is based on the army's classification of health and physical capabilities for different forms of service, will be used to analyze changes in the social composition of the army. A military career may include numerous positions if a man served in the army for several years, which was common in the Finnish Army of World War II. On average, a man who fought in the war has 7.6 different positions and 5.7 different units in his career. Half of these men have 10 or more records with changes in the variables position, military unit, or task.

The table "Experiences" contains two types of data (see Table 5). First, it includes undated military career information like the number of battles men fought in, training for military task, and evaluations. Second, it contains the variables we constructed from the two other main tables, summarizing soldiers' military careers and experiences like the duration of wartime service and participation in different stages of the war. These variables cover our prime targets of inquiry and are created at this point to simplify their query during analyses.

The main tables are linked to three other datasets. The most important of these is the database of soldiers killed in the war constructed by the National Archive of Finland. All soldiers in our sample who died in the war can be linked to this database because they were picked from it. This dataset offers additional information regarding soldiers' wartime deaths, such as the location of their mortal wounding and death and its manner (died in action/hospital/went missing/etc.). The table "Locations" provides additional information about the places of birth, residence and death recorded at municipality level in the database. The table includes their wartime population figure, administrative level (municipality, market town, city), province and coordinates. Due to the size of our sample, it is necessary to convert places to provincial level in regional analyses. Lastly, the occupation titles will be linked to the HISCO classification.

Table 5 *Military experience data in FA2W database and constructed variables*

Variable	Value
Number of Battles	Number
Military Branch Education	Finnish Army Branches
Military Task Education	Definition
Evaluation (Punctuality, Diligence, Powers of Observation, Military Development, Personal Conduct)	Good/Average/Poor
Membership of Voluntary Militia	Yes/No
Survival	Survived/Killed/Did not Participate
Participation in wars (Winter War, Continuation War, Lapland War)	Yes/No
Military Ranks Classes	Rank and File, NCO, Officer
Duration of Conscription	Number of Days
Duration of Wartime Service	Number of Days
Number of Mobilizations	Number

The FA2W database offers numerous angles from which to study the effects of war, but there are also limitations and challenges inherent in its data. Some variables are affected by recording practices that cannot be elaborated here. However, two clear issues are worth mentioning. Firstly, the data becomes more comprehensive as it becomes more recent. During the 1920s, the first decade of the formation of the Finnish Army, recordkeeping was less systematic and the form of the military service record was less detailed than its later versions. From the 1930s onwards the military records are more comprehensive due to the introduction of a new form and guidelines. For the World War II years, 1939–1945, the data is available in its greatest detail because numerous different documents were used during this active period of service.

Secondly, socio-economic information, such as occupation, marital status, and education, was gathered at two time points: at conscription and during the war years. The military service record form in use from the early 1930s to the end of the war has two sections: peacetime and wartime entries, which were the same for men who joined in wartime. The issues with this information concern the updates of the records and the form. The military service record form was renewed in 1930 and again in 1945, and in both cases information was copied from old forms onto new ones if a man belonged to the reserve. This was not a problem when the army archived old records. However, especially during the upgrade of 1945, some military districts destroyed records from the pre-war and war years after copying information onto new forms until an order was issued for their preservation (Ylönen-Peltonen, 2021). Another challenge is updates during refresher training before or after the war. If a man joined these events, conscription and wartime socio-economic information was sometimes updated to reflect their later status.

These recordkeeping practices mean that information on socio-economic status from both conscription service and the war years is not always accurate, even if it was originally recorded in the military service records. This problem can be overcome in some cases with other documents, such as medical histories, which contain information on place of residence and occupations during the conscription and war periods. However, these recordkeeping practices mean that even if we have occupation information about 93% of the sample and 98% of men who served in the war, we do not have this information in the same detail for both periods. The conscription period data is particularly deficient, and occupation information is missing from one third of the sample. Some of this data can be collected later from the draft lists.

6 DATA ENTRY

The construction of the database began on funding granted by the Finnish Cultural Foundation for the STASKO project in 2017 with the designing of the initial sample. In 2018, over 60,000 pages of documents were photographed. To enter the data, we use the web-based database management platform REDCap (<https://projectredcap.org>, see also Harris et al., 2009; Harris et al., 2019).

Between 2018 and 2020, salaried research assistants and the author of this article entered one quarter of the data into the database. Since the records include sensitive medical information on people who in rare cases could still be alive, we could not make use of crowdsourcing. Furthermore, crowdsourcing would have been difficult to implement due to the variety and complexity of the data. The military service record files contain numerous different types of documents where recording practices changed over time, meaning that research assistants need extensive training on the material and its historical context to be able to enter the data. The data entry process is relatively time-consuming. With over 60 variables to be collected from an average number of 14 documents, it takes on average 20–30 minutes to enter one man's information into the database. The laboriousness of this work, which is largely due to the richness of the military service record files, was something of a surprise for us and delayed the completion of the database.

After one quarter of the data had been entered, analysis of this data revealed distortions in the sample as described above. These problems were investigated in 2020, and the new stratified sample was designed in 2021. The necessary additional records have since been collected and data entry is slowly progressing. Additional funding permitting, the first phase of the database described in this paper is scheduled for completion in 2024.

Figure 6 A segment of REDCap data input form

7 RESEARCH OPTIONS AND FUTURE PLANS

The FA2W database offers many opportunities for both population-level demographic history and life course analysis. A central goal for the database is a better understanding of the impact of World War II on different sections of Finnish society. So far, this question has received relatively little scholarly and public attention. This may be due to the importance of World War II for Finnish national unity, but lack of data has also played a role. The available data has already generated distributions of Finnish war casualties by regions, age cohorts, and social classes. However, few studies have analyzed issues like mortality rates in specific social groups due to the lack of comprehensive data on surviving soldiers. These studies show that these are important issues: there were wide regional differences in mortality rates and the most severely affected seem to have been the rural poor (Kivimäki, 2019; Toivonen, 1998; Waris, 1948).

As a representative sample of all Finnish men of the age cohorts who fought in the war, the FA2W database will make it possible to investigate the direct consequences of war, such as mortality, injuries, and sickness rates among social classes, age cohorts, linguistic groups, and regions. It will enable us to connect these differences with Finnish conscription training and mobilization practices, such as how social groups were trained for different military tasks, who was sent to the trenches, and who could stay home, which crucially determined the distribution of the wartime burden. The FA2W database offers a base to study these inequalities in military sacrifice during the period of total warfare of the World Wars. This issue has been studied previously particularly in the US Army in the post-World War II period (Barnett, Stanley, & Shore, 1992; Kriner & Shen, 2010; Merli, 2000), in which lower classes and minorities have carried the heaviest burden of wars, but it has not gained similar attention in European nations involved in the World Wars that presumably fought total, shared warfare with universally conscripted armies. As nation-states also habitually frame their wars as shared endeavors, the unequal consequences of wars have not been thoroughly publicly acknowledged, and as Kriner and Shen (2016, p. 554) note, it is a topic that "academic scholarship has not seriously explored".

For life course analysis, the main value of the database is that it permits detailed examination of how different forms of wartime service affected soldiers' mortality and survival in war. This can be done with various data such as the duration that the men spent in frontline troops, military branches, military units, and military tasks. This analysis can be further enriched with contextual information about casualty rates in these positions at different phases of the war. We can, for example, calculate how many weeks the sampled men stayed in heavy frontline infantry combat, stationary warfare duty, or rearguard tasks and whether the periods of combat were defensive or offensive in nature. This makes it possible to examine the relationship between mortality and the nature of soldiers' military service and exposure to combat and violence in a very detailed manner. Many military science studies have shown that, for example, infantrymen die in wars more often than artillery personnel (Bellamy, 2000; Buzzell & Preston, 2007), but the FA2W database also enables us to study the forms of service inside the military branches.

The data on wartime service and experiences are also central in the planned future of the database and the research it enables. When the initial work on soldiers' prewar and wartime lives is done, the database will be enriched with postwar data to study the long-term effects of war and exposure to violence. This data will include information about causes of death, which is already partially available in the military records, and information on socio-economic status and health from nationwide records and registers that became available after the war. One plan is to connect the sample to the census records collected every 10 years from 1950 onwards. This will provide immense opportunities to investigate the long-term effects of war and violence. In particular, it will be possible to investigate how much long-term life chances were tied to degree of exposure to violence and combat and to compare the effects of experiencing death face to face in the trenches to service in relatively safe circumstances in auxiliary duties behind the lines. This may elucidate the effects of violence more clearly than studies based only on aggregate data like wartime age cohorts or soldiers of wars in general (Saarela & Finnäs, 2012).

In general, the Finnish case will offer an interesting chance to scrutinize the long-term effects of war. As MacLean and Elder (2007) state in their review of research on the consequences of military service in veterans' post-war lives, it is evident that veterans' fates have varied considerably in different historical circumstances. A comparison between countries like the United States, where soldiers have been drafted mostly from the lower classes, and Finland, where the experience of World War II was shared by all social strata in a spirited and united manner, could offer interesting results. It has been suggested, for example, that PTSD was a lesser problem among Finnish veterans of World War II because this war has been very important in Finnish culture (Hautamäki & Coleman, 2001).

Furthermore, as the database offers a representative picture of the Finnish male generation of the early 20th century, it can also be used to examine questions about the social structure and health of Finnish society and the life courses of the war generation throughout the century. One of the most promising research possibilities is examining connections between health during the conscription period and later life. The military records include rich medical data, which can be added to the database in support of this research.

8 CONCLUSION

In this article, I have introduced the first steps of a database that was born out of curiosity while conducting research on other nearby topics. As this is the first demographic database built by our research group, the process has been a learning experience with constant re-evaluation of our methods and decisions. The foundation of the database is now in place, but many changes are still sure to come when we head into the finalization of its first phase. While constructing a large database has been a laborious task, we believe it to be a worthy undertaking because it promises to provide invaluable empirically grounded insight into human experiences of war, a topic that our research group among many others has examined in recent times, mostly from a cultural historical perspective. When the groundwork is done, we envisage that the database will serve as a foundation to investigate these issues for many years to come.

ACKNOWLEDGEMENTS

The construction of the Finnish Army in World War II Database began thanks to a research grant from the Finnish Cultural Foundation and the Harry Hendunen Fund. The writing of this article was supported by the Society of Swedish Literature in Finland. The construction of the database is led by Ville Kivimäki and the database designed by Ilari Taskinen in Tampere University. Katariina Eskola, Siiri Simppanen, Vikke Niemi, and Miiko Siivonen have worked as research assistants on the project. Raija Ylönen-Peltonen and others in the National Archives of Finland have been truly helpful in assisting us with the military service record collection. Turkka Näppilä deserves thanks for his help with REDCap, Jyrki Ollikainen with his advice on sample design and Jarmo Peltola for sharing his expertise on population statistics and data. I want to further thank Ilkka Jokipii, Joonas Kumpulainen, and Rick Mourits for their advice on different matters of the project and lastly, the Radboud Group for Historical Demography and Family History for hosting my research visit and work with the database.

REFERENCES

- Ahlbäck, A. (2014). *Manhood and the making of the military. Conscription, military service and masculinity in Finland, 1917–39*. Farnham: Ashgate.
- Ahoranta, T., & Ortamo-Närvä, A.-M. (2012). Arkisto säilyttää sankarivainajien muiston [Archive preserves memory of fallen heroes]. *Sukutieto*, 29(3), 12–13.
- Barnett, A., Stanley, T., & Shore, M. (1992). America's Vietnam casualties: Victims of a class war? *Operations Research*, 40(5), 856–866.
- Bellamy, R. F. (2000). Why is marine combat mortality less than that of the army? *Military Medicine*, 165(5), 362–367. doi: [10.1093/milmed/165.5.362](https://doi.org/10.1093/milmed/165.5.362)
- Bessel, R. (2015). Death and survival in the Second World War. In M. Geyer & A. Tooze (Eds.), *The Cambridge history of the Second World War. Volume III, Total war: Economy, society and culture* (pp. 252–276). Cambridge: Cambridge University Press. doi: [10.1017/CHO9781139626859.012](https://doi.org/10.1017/CHO9781139626859.012)
- Buzzell, E., & Preston, S. H. (2007). Mortality of American troops in the Iraq War. *Population and Development Review*, 33(3), 555–566. doi: [10.1111/j.1728-4457.2007.00185.x](https://doi.org/10.1111/j.1728-4457.2007.00185.x)
- Fogel, R. W., Costa, D. L., Haines, M., Lee, C., Nguyen, L., Pope, C.,... Yetter, N. (2000). *Aging of veterans of the Union Army: Version M-5*. Chicago: Center for Population Economics, University of Chicago Graduate School of Business, Department of Economics, Brigham Young University, and The National Bureau of Economic Research. Retrieved from <https://www.nber.org/programs-projects/projects-and-centers/union-army-data/union-army-data-citation-and-use-early-indicators-data>
- Fornasin, A., Breschi, M., & Manfredini, M. (2019). Deaths and survivors in war: The Italian soldiers in WW1. *Demographic Research*, 40(22), 599–626. doi: [10.4054/DemRes.2019.40.22](https://doi.org/10.4054/DemRes.2019.40.22)
- Harris, P. A., Taylor, R., Minor, B. L., Elliot, V., Fernandez, M., O'Neal, L.,... REDCap Consortium (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, 95. doi: [10.1016/j.jbi.2019.103208](https://doi.org/10.1016/j.jbi.2019.103208)

- Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Gonde, J. G. (2009). Research electronic data capture (REDCap): A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, 42(2), 377–381. doi: [10.1016/j.jbi.2008.08.010](https://doi.org/10.1016/j.jbi.2008.08.010)
- Hautamäki, A., & Coleman, P. G. (2001). Explanation for low prevalence of PTSD among older Finnish war veterans: Social solidarity and continued significance given to wartime sufferings. *Aging & Mental Health*, 5(2), 165–174. doi: [10.1080/13607860120038348](https://doi.org/10.1080/13607860120038348)
- Karjalainen, M. (2014). Tietojen vertailua, korjaamista ja asiakaspalvelua: Suomen sodissa 1939–1945 menehtyneiden tietokanta [Comparing information, corrections and customer service: Database of fallen in the Finnish wars 1939–1945]. In P. Happonen (Ed.), *Kleion pauloissa* [In Kleio's spell] (pp. 277–283). Helsinki: SKS.
- Kempainen, I. (2006). *Isänmaan uhrin: Sankarikuolema Suomessa toisen maailmansodan aikana* [The nation's heroes. Military death in Finland during the Second World War]. Helsinki: SKS.
- Kinnunen, T., & Kivimäki, V. (Eds.). (2012). *Finland in World War II: History, memory, interpretations*. Leiden & Boston: Brill.
- Kivimäki, V. (2013). *Battled nerves: Finnish soldiers' war experiences, trauma, and military psychiatry, 1941–44* (Doctoral dissertation). Åbo Akademi University, Turku. Retrieved from <http://www.doria.fi/handle/10024/90586>
- Kivimäki, V. (2019). Sankariuhri ja kansakunta: Suomalaiset sotakuolemat 1939–1945 [Fallen heroes and nation: Finnish war deaths 1939–1945]. In I. Pajari, J. Jalonen, R. Miettinen & K. Kanerva (Eds.), *Suomalaisen kuoleman historia* [History of Finnish death] (pp. 277–310). Helsinki: Gaudeamus.
- Kriner, D. L., & Shen, F. X. (2010). *The casualty gap: The causes and consequences of American wartime inequalities*. Oxford: Oxford University Press. doi: [10.1093/acprof:oso/9780195390964.001.0001](https://doi.org/10.1093/acprof:oso/9780195390964.001.0001)
- Kriner, D. L., & Shen, F. X. (2016). Invisible inequality: The two Americas of military sacrifice. *University of Memphis Law Review*, 46(3), 545–635. Retrieved from <https://www.memphis.edu/law/documents/kriner-shen46.pdf>
- Kronlund, J. (1988). *Puolustusvoimien rauhan ajan historia. Suomen puolustuslaitos 1918–1939* [Peacetime history of defence forces. Finnish defence establishment 1918–1939]. Porvoo: WSOY.
- Kurenmaa, P., & Lentilä, R. (2005). Sodan tappiot [Casualties of war]. In J. Leskinen & A. Juutilainen (Eds.), *Jatkosodan pikkujättiläinen* [Small giant of Continuation War] (pp. 1150–1162). Helsinki: WSOY.
- MacLean, A., & Elder, G. H. (2007). Military service in the life course. *Annual Review of Sociology*, 33, 175–196. doi: [10.1146/annurev.soc.33.040406.131710](https://doi.org/10.1146/annurev.soc.33.040406.131710)
- McCalman, J., Kippen, R., McMeeken, J., Hopper, J., & Reade, M. (2019). Early results from the 'Diggers to Veterans' longitudinal study of Australian men who served in the First World War: short- and long-term mortality of early enlistees. *Historical Life Course Studies*, 8, 52–72. doi: [10.51964/hlcs9307](https://doi.org/10.51964/hlcs9307)
- Merli, M. G. (2000). Socioeconomic background and war mortality during Vietnam's wars. *Demography*, 37(1), 1–15. doi: [10.2307/2648092](https://doi.org/10.2307/2648092)
- Modell, J., & Haggerty, T. (1991). The social impact of war. *Annual Review of Sociology*, 17, 205–224. doi: [10.1146/annurev.so.17.080191.001225](https://doi.org/10.1146/annurev.so.17.080191.001225)
- Nurminen, T. (2008). Muuttuva armeija [Changing army]. In J. Kulomaa & J. Nieminen (Eds.), *Teloitettu totuus — Kesä 1944* [Executed truth — Summer 1944] (pp. 47–75). Helsinki: Ajatus Kirjat.
- Saarela, J., & Finnäs, F. (2012). Long-term mortality of war cohorts: The case of Finland. *European Journal of Population*, 28(1), 1–15. doi: [10.1007/s10680-011-9246-x](https://doi.org/10.1007/s10680-011-9246-x)
- Stead, W. W., Hammond, W. E., & Straube, M. J. (1982). A chartless record – Is it adequate? *Proceedings of the Annual Symposium on Computer Application in Medical Care*, Nov 2, 89–94. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2580254/>
- Toivonen, T. (1998). War and equality: The social background of the victims of the Finnish Winter War. *Journal of Peace Research*, 35(4), 471–482. doi: [10.1177/0022343398035004](https://doi.org/10.1177/0022343398035004)
- Urlanis, B. Ts. (1971). *Wars and population* (L. Lempert, Trans.). Moscow: Progress Publishers.
- Waris, H. (1948). *Suomalaisen yhteiskunnan rakenne* [Structure of Finnish society]. Helsinki: Otava.
- Winter, J. M. (1986). *The Great War and the British people*. London: Macmillan.
- Ylönen-Peltonen, R. (2021). Sotapolun jäljillä arkistossa: Talvi- ja jatkosodan henkilöhistoriallisista arkistolähteistä Kansallisarkistossa [Tracing soldier's war path: Personal history records from Winter War and Continuation War years in national archive]. *Sotahistoriallinen aikakauskirja*, 41, 201–219.

Twenty-three major databases containing historical longitudinal population data are presented and discussed in this edited volume, focusing on their aims, content, design, and structure. Some of these databases are based on pure longitudinal sources, such as population registers that continuously observe and record demographic events, including migration and family and household composition. Other databases are family reconstructions, based on civil records. The third and last category consists of semi-longitudinal databases, that combine, for instance, civil records and censuses and/ or tax registers. The volume traces the origins of historical longitudinal databases from the 1970s and discusses their expansion worldwide, in terms of sources and hard- and software. The contributions highlight the unique genesis and common developmental arcs of these databases, which are rooted in the fields of quantitative history, social and demographic history, and the history of ordinary people. The importance of these databases in advancing knowledge and insights in various disciplines is emphasized and demonstrated, along with the challenges and opportunities they face.

The collection of technical descriptions of these databases represents the most comprehensive and up-to-date overview of large databases with longitudinal micro-data on historical populations. It includes descriptions of databases from Europe, North America, East Asia, Australia, South Africa, and Suriname. Technical details, in terms of data entry, cleaning, standardization, and record linkage are meticulously documented. The volume is a must-have for all scholars in the field of historical life course studies.

Kees Mandemakers is affiliated member of the Radboud Group for Historical Demography and Family History. He is senior fellow at the International Institute of Social History and Emeritus Professor of Large Historical Databases at the Erasmus School of History, Culture and Communication of Erasmus University Rotterdam.

George Alter is Research Professor Emeritus at the Institute for Social Research at the University of Michigan.

Hélène Vézina is professor in the Department of Human and Social Sciences at the Université du Québec à Chicoutimi.

Paul Puschmann is Assistant Professor of Economic, Social and Demographic History at Radboud University Nijmegen.

ISBN 978-94-9329-617-6



9 789493 296176 >

Radboud Universiteit



www.radbouduniversitypress.nl