THEORY AND PRACTICE OF HISTORICAL CENSUS DATA HARMONIZATION

THE DUTCH HISTORICAL CENSUS USE CASE:
A FLEXIBLE, STRUCTURED AND ACCOUNTABLE APPROACH USING LINKED DATA

This interdisciplinary research was conducted in the context of the CEDAR (Census Data Research) project which was part of the Computational Humanities programme, of the KNAW E-humanities Group in Amsterdam. In this collaboration the International Institute of Social History (IISH), Erasmus University Rotterdam, Data Archiving and Networked Services (DANS), Radboud Universiteit Nijmegen and the Vrije Universiteit in Amsterdam (VU) worked closely together.

Theory and Practice of Historical Census Data Harmonization
The Dutch historical census use case: a flexible, structured and accountable approach using Linked Data technology.

Theorie en praktijk van historische volkstellingen harmonisatie
De Nederlandse historische volkstellingen: een flexibele, gestructureerde en verantwoordelijke benadering met behulp van Linked Data technologie.

Proefschrift

ter verkrijging van de graad van doctor aan de Erasmus Universiteit Rotterdam

Op gezag van de
rector magnificus

Prof. Dr. R.C.M.E. Engels

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op

donderdag 17 januari 2019 om 13:30 uur

Ashkan Ashkpour
geboren te *Tehran, Iran*

Erasmus University Rotterdam

Promotiecommissie

Promotor*:*

Prof. dr. C.A. Mandemakers

Overige leden:

Prof. dr. J. Kok

Prof. dr. F.M.G. de Jong

Prof. dr. C.M.J.M. van den Heuvel

Copromotor:

Dr. O. Boonstra

PREFACE

Writing this dissertation and the entire journey that came with it has been the most enriching experience I had. Being at the forefront of digital humanities research in the Netherlands and having the pleasure to work in an interdisciplinary project where we combined (as one of the first) Linked Data technologies with social historical research has shaped my interest and passion for this field, for many years to come. The knowledge and experienced I gained in order to grow as a researcher and person is something which I'm very thankful for. Whether it was the topic and dataset which I became to love or the unexplored terrains we were exploring, it was always the people who made it a pleasure.

First, I would like to thank the people directly involved in my research. My sincerest gratitude goes to my supervisors Kees Mandemakers and Onno Boonstra. Kees, it was my pleasure to be your student for the last years and I truly appreciate all the lessons and wisdom you shared with me. Whether it was your attention for detail and critical thinking during our meetings, the joint writing sessions, the drinks, food and good times we shared in our many trips together or the personal talks, I will carry these moments with me for always. Onno, our distance prevented us to meet more frequently but your feedback during every stage of the process was always very enlightening and practical to me. Your work has been extremely valuable and inspiring throughout the project and has been used as the groundwork in many of my research endeavors.

Next I would like to thank the members of the PhD committee for devoting their valuable time and evaluating my dissertation,

Franciska de Jong, Jan Kok, Charles van den Heuvel, Hein Klemann, Peter Doorn and Karin Hofmeester, it's an honor to be evaluated by such an impressive committee.

Throughout the year I had the pleasure of being affiliated with several institutes and working with researchers from various domains. One of the institutes which was greatly involved in the digitization of the historical censuses and highly involved in the PhD project was DANS. Being one of the original caretakers of the census (even to this date) and initiators of the CEDAR project, the work done by DANS and its wonderful people is of high value for researchers across various domains. First, I would like to thank Andrea Scharnhorst, our project leader and binding force in CEDAR. Thank you for your tireless efforts of bringing and holding together such an interdisciplinary project and making me feel at home from the very beginning. Peter Doorn, your experience and passion for the subject over the past decades is very contagious. Thank you for being an advocate for such a valuable data source, for such a long time. I also would like to thank the wonderful colleagues of the DANS research group.

The e-humanities group was truly a unique, and trailblazing, initiative which gave many of the PhD students involved in various cross disciplinary projects a very strong basis and network to build on and thrive in the current landscape of digital humanities research projects. Thank you Sally Wyatt, Andrea Scharnhorst, Jeannette Haagsma and Anja de Haas for this unique experience. Sally, whether it was one of the many talks you gave, the stories you shared or your vision and love for this field, you were always very inspiring.

Even though I have shared many workplaces, my home base has always been the IISH, for the last six years. Being at the source of knowledge and expertise in my field made me feel very honored to work at this institute, also beyond the CEDAR project.

During the PhD project I have had the pleasure to work with and get to know many colleagues, however there is only one which I call 'Mi Hombre'. Albert, you were the friend I hadn't met yet. You were my first colleague in this journey and became a friend for life. We shared many great and significant moments together and just like our PhD time, I'm certain we will share many more great moments to come.

Finally, I would like to give a special thanks to my family. Father, Mother, thank you for your unconditional support, sacrifices and giving us the opportunity to reach for the sky. Saman, Fereshteh and the little princess Ramona, it's been wonderful to see your beautiful family grow. Becoming an uncle and the joy it gave me during this process was the greatest gift of all.

CONTENTS

TABLE OF FIGURES

TABLE INDEX

1. INTRODUCTION

1.1 SUBJECT OF THIS STUDY

Censuses contain a wealth of information about nations and societies. They structurally capture societal information needs at given times in the past. Throughout history, the censuses have served to provide information to governments, i.e. to understand the development of the nation and its population on several fronts, for decision-making purposes. The historical censuses can currently still mean a lot for researchers. Historical censuses are one of the scarce, reliable and large-scale statistical data sources we have about our nation's past. They often are the only comprehensive statistical datasets with regards to the demographic and socio-economic life of our past. They are large scale as the census covers the entire population and geographical context of a nation (from the biggest city to the smallest village). Furthermore, they are considered as one of the most reliable sources as censuses are taken consistently at regular intervals and conducted in a well prepared manner by governments. However, looking back at our history through the census has proven to be a challenging endeavor. With all its positive traits, the use of historical census data for longitudinal research purposes has been hampered by the lack of comparability over the years which resulted in less use of this valuable data.

Throughout history, the use and public opinion of the censuses have changed significantly. Censuses were first primarily used as a tool for taxation or war purposes and mostly regarded as a

'suspicious thing' by those being enumerated. In the course of the nineteenth century we see a shift to its acceptance by the public and nowadays as a tool for governmental decision making and a valuable resource to answer pressing societal demands to improve the quality of life (Daniels 2004). The example below shows a part of an article in a newspaper announcing the U.S census for 1900 (Hepps 2015) with a rather obligatory connotation:

> *"Don't lie. When the census enumerator comes around June 1 tell him the truth. If you don't you will go to the bad place and if he finds out you may go to a worse place….[]…Some of the questions the enumerators are expected to ask may seem a little obnoxious, but that is not the fault of the enumerator. He is there to ask all the questions as printed, and he is expected to get true and correct replies. If any person refuses to answer them, he is liable to arrest, fine and imprisonment."*
> *(Hepps 2015, para. 3)*

Over the years a new goal was added to the practical uses of the census. Next to being used as a tool for governmental decision making, the census has become a valuable resource for research. The potential of historical census data for a variety of users such as social scientist, historians, socio-economic historians, demographers, archivists, students, governments and general public etc. is far from being exhausted (Higgs, 1996; Ruggles and Menard 1995; Doorn and Maarseveen 2007). However, the challenges faced when using historical census data in its original form has almost discouraged researchers to the point of neglecting

the census as a valuable resource. For example, in his article 'The census and the historical demographer', Doorn (2012, p. 30) presents the pressing question: *"Is the role of censuses for historical demographers […] over? The census seems to have become less en vogue as a source of demographic research*". However, topics such as industrial restructuring, migration, aging of the population and financial crises in a world of accelerated change are still very current topics in Europe. Learning from our past through the census allows us to understand the interrelation between macro-economic change, policy changes, demographical shifts, labor markets, communities, national wealth and much more. However, the data needed to answer these questions are difficult to produce given the scatteredness and dissimilarity of the censuses over the years.

In order to use the Dutch historical censuses in a longitudinal and comparative way researchers are often confronted with the need of integrating the dissimilar structures, variables, values and classification systems, before they can use the data in a uniform way across time and space. The various solutions regularly used by some historians to deal with these integration problems are often loosely referred to as harmonization. This study contributes to the advancement and curation of the Dutch historical census data, and its use by the community of social and economic scholars, historians, and beyond. More specifically, this research focuses on the theory and practice of aggregate historical census data *harmonization* over time and space. The realization of a fully integrated census dataset will give a boost to the use of such data by researchers. The harmonization challenges presented by historical censuses are one of the most notorious ones and often also present, in one way or the other, in other historical datasets.

By addressing these challenges we provide generic solutions for the harmonization of aggregate statistical sources in general. In order to achieve this, we explore the possibilities provided by the Resource Description Framework (RDF) and Linked Data principles. We do this for both methodological and practical solutions. Modeling the aggregate Dutch historical census data across time will provide a workflow, methods, tools, ontologies and more for other researchers to work with and will offer clear cross-disciplinary benefits.

In this chapter we continue with the description of the Dutch historical censuses (1.2) and the wealth of information it contains (1.2.1). In section 1.2.2 we look at several key historical comparative studies using the census and its potential for research. In section 1.3 we look at the main problems of the historical censuses, hindering the use of this valuable dataset for research over time and space. The goal, our contributions and the research question of this study are explained in section 1.4, followed in section 1.5 by the context in which this study was performed, i.e. the CEDAR project. Section 1.6 of this chapter provides a detailed description of the content of this study. It consists of the different sections, chapters and various sub-research questions which are answered in each section of this study. We close this chapter with an overview of shared work of the publications that are used in this dissertation (1.7).

1.2 THE WEALTH AND VALUE OF THE DUTCH HISTORICAL CENSUSES

An important aspect of the historical censuses is their potential to study social and economic change over long periods of time. They provide information about housing needs and valuable socio-economic data such as occupations. And, of course, as a source for demographic information about nations, the census is an irreplaceable asset. Whether we are interested in answering very specific questions about small geographical areas and sub-populations or more general questions about the development of populations in different provinces or states, the census often remains the only source to find the necessary data.

1.2.1 BACKGROUND

The first general enumeration in the Netherlands took place in the Batavian Republic, in 1795. It paved the way for the first official census in the Netherlands, held in 1829. From this year onwards the census was held every ten years until 1971, except 1940 and 1950 which were replaced by 1947 due to the Second World War. Censuses taken during this period in the Netherlands are called the historical censuses. They distinguish themselves from the modern census in the way the population was enumerated, i.e. by going door to door and collecting the information by hand. Due to more concerns and protest of the public with regard to privacy issues, political but also budgetary aspects, 1971 was the last door to door census (Den Dulk and Van Maarseveen 1999). From 2000 onwards, the electronic municipal population registers are used to collect the census data. However manual enumeration still

occurs in other countries, especially in the Anglo-Saxon countries as a consequence of the lack of population registers. Through these extensive, time and money consuming enumeration of the population, the historical censuses have become one of richest sources to study our past on a large scale.

When referring to the Dutch historical censuses we distinguish three different forms, all collecting information on different aspects in society. These are the 'Population', 'Occupation' and 'Housing' census. The 'population census' is one of the largest historical demographical sources of our past. It contains information about the population at given times, with regard to characteristics such as age, gender, marital status and religion. The increasing demand for information about the occupational structure and its developments led to the introduction of the occupational census in 1849. Information collected in the occupational census was used to study the development of the occupational structure in the Netherlands on various geographical levels (De Jonge 1966, Van Dijk and Verstegen 1988). We could for example study the growth and decline of specific occupations due to specialization or differentiation. Moreover, occupation is one of the few variables which provides insights in an individual's relation to society in a distinct way. From 1889 onwards, the Dutch occupational census even used a classification to distinguish occupations into four groups of social positions, allowing us to identify whether an inhabitant who was for example counted as 'watchmaker' actually was a production worker / craftsman, a foreman, managerial function or the owner of a small or large company. The third census is that of the housing census. This census has played an important role in decision and policy making with regard to the housing situation. For example, after

the Second World War the housing census of 1956 was used to gather data about the housing market in order to deal with the problem of housing shortage (Van Maarseveen, 2002). The housing census contains information about the size and structure of the housing stock, the housing needs, reserves etc. The level of information found in the housing census is very detailed. Besides standard questions which were asked in all housing censuses such as the number of people living together and the number of rooms they shared, the housing censuses also introduced the so-called 'morality questions' (zedelijkheidsvragen). Questions such as the number of box beds and the frequency of co-sleeping siblings until a certain age in the many one- or two-room apartments were a prominent part of the historical housing census. These phenomena were thought to be a threat to public health and such questions with a moral background were therefore used throughout the housing censuses of 1909-1947 (Van der Bie, 2007).

Efforts to provide greater access to the Dutch historical census data started almost two decades ago in 1997. The first step in this process was to preserve and provide better access to the data by scanning the original books. In total 193 books consisting of 43,000 pages were digitized during this process. Tens of thousands of images were consequently created and made available via various websites, cd-roms, archives etc. Although a great improvement compared to physical access to the books often found in libraries, the images are extremely difficult to handle. Therefore, after this period the focus shifted towards content conversion and the images were (manually) transcribed into Excel tables. During this process, the choice was made to represent the images as one to one copies in Excel. This means that both the

data as well as the structure / layout of the tables were copied into Excel in a strict source-oriented manner. In total this resulted in 2249 separate Excel tables. These tables are the point of take-off in this study.

1.2.2 THE (RE)USE OF THE DUTCH HISTORICAL CENSUSES

The Dutch historical census is one of the most used statistical datasets in the Netherlands by historians who study the nineteenth and twentieth century. The potential of the historical censuses for research purposes has shown some interesting uses by researchers thus far. Interestingly, we also find studies where the census is used in combination with other datasets to answer questions that span outside the realm of censuses. In order to convey the richness and potential for research of the census we present several interesting studies which use or build primarily on the Dutch historical censuses in this section.

Since the start of the digitization the census has become much better accessible and it has been used by many researchers. The census is used to study topics such as the development of the population in general, development of various characteristics related to the population (e.g. size, marital status, age etc.), the structure of employment, occupational development, church and religion, housing and migration etc. In order to show the variety of subjects and richness of the census we identify three main areas in which the census excels as a valuable source for comparative research, i.e.: demographic studies, socio-historical studies and studies which focus on economic aspects.

An early and significant (comparative) study using census data is that of Van Dijk and Verstegen (1988). In their work called "Dienstverlening in Nederland en Duitsland, tussen eerste wereldoorlog en welvaartsstaat", Van Dijk and Verstegen looked at several societal changes in industrialiazed countries across time. The data used in their work is extracted from the occupational censuses of Germany and the Netherlands (1880-1980). The development of the occupational statistics in both countries is given primary attention. Some of the key topics addressed in their study are the rise of the 'service sector', the shift from traditional to modern service occupations and of the female participation in these sectors.

At the turn of the century, in the year 2000, together with the celebration of its 100th anniversary the Dutch statistics bureau (CBS) published a book "Nederland een eeuw geleden geteld" (Van Maarseveen and Doorn 2001). In this book thirteen different studies are presented, which primarily make use of the most elaborate censuses ever held in the Netherlands, i.e. the 1899 census. A variety of studies are presented on topics such as the changing population structure, the growth of the population (Van Poppel 2001), and analysis of the foreign (migrant) population according to their origin, gender distribution and occupational structure (Van Eijl and Lucassen 2001). Studies focusing on social aspects of society and the population are also well represented. In his study Noordam (2001), based on the influential ideas of Edward Shorter, looks at the modern family and what it entails for the Netherlands at the turn of the century. Using the census he finds that the civilization around 1899 was moving towards a society with much stricter moral standards. In fact, the study showed that the Netherlands, compared with the rest of Western

Europe, had the lowest number of extramarital births, divorces and a very low number of forced marriages, a relative high age of marriage etc. In another study on societal aspects, Knippenberg (2001) focused on secularization and the segregation of the society into different religious denominations, contributing to our current knowledge on the changing population composition throughout history. Next to demographic and sociological studies the census is also a valuable source for the study of economic aspects of societies in the past. Horlings (2001) studies topics such as employment and economic modernization and the structure of the labor force using the historical census.

The studies mentioned above focused on the most detailed census, i.e. that of 1899, and are examples of the potential of the data. However we also have contributions using other years and even studies comparing censuses over time. In corporation with the CBS (the Netherlands Statistics Bureau) DANS published a book called "Twee eeuwen Nederland geteld: onderzoek met de digitale Volks-, Beroeps- en Woningtelligen 1795-2001" (Boonstra et. al 2007). The topics presented in this book range from migration, ageing, fertility, household, economic development, social relations, geography, housing situation, religion, entrepreneurship and much more. The value of the census is most recognizable in its use for longitudinal studies. We find several studies spanning over time which use the census as a key data source or use it to provide context. For example, in their study of the foreign migration in the Netherlands between 1795 and 2006, Nicolaas and Sprangers (2006) look at the impact of migration on the population composition for over two centuries. Interestingly, in this study the census is used in combination with other datasets. Another fascinating topic of study is the employment rate of

women above the age of 50 (Oudhof and Boelens 2006) between 1849 and 2006. In this research the census played an important role in determining the development of the labor force across time. Other longitudinal studies can be found on topics such as infrastructural development, studied by Groote en Tassenaar (2006) for the provinces of Groningen and Drenthe between 1820 and 1915. In this research the census is used to study the distribution of the population on the level of neighborhoods. The geographical variables of the census provide many opportunities to link the census with spatial data from other sources. Doing so we can study change on various geographical levels over time as well as space.

It will be clear by now that the importance, richness and variety of research questions that can be answered using the Dutch historical census is unmistakable. These various studies are based on census data made available after the various digitization projects. Although these efforts gave the census a new stimulus, its true potential to study changes over time still had not been reached. Only ten of the thirty studies published in these books use the census for longitudinal studies. To make the data comparable these researchers have put extensive efforts in data cleaning, correcting, mapping, standardizing etc. Unfortunately their decisions, corrections, standardizations and other time consuming activities are not (easily) reusable by others as they are not archived in a systematic and reproducible way.

So although the census is one of the most comprehensive and frequent used historical statistical datasets, it is definitely not one of the easiest to use for longitudinal analysis. This is however not due to the lack of interest in other years by researchers, but mostly

due to major changes the censuses faced from one year to another. As a result most studies and projects working with the historical census data focus on a single year or a series of census years in which the census had not changed significantly.

1.3 PROBLEMS HAMPERING THE USE OF THE DATA

One would expect that after the digitization wave of the Dutch historical censuses, the use and recognition of the census as a valuable research asset would increase and contribute to more longitudinal studies. However, decades after the first digitization efforts started, we find that this is not the case yet. In practice this has resulted in researchers using only isolated sections or parts of the census which are more easily comparable (Van Maarseveen 2008), which is particularly the case with the Dutch census data.

The possibilities to use the historical censuses by the scientific community is severely limited by the unconnectedness of the data, due to the heterogeneity in structures, variables and classifications that are used. Consequently, researchers tend to seek their own specific solutions which are only justifiable by their interpretations and not their actions. This results in non-repeatable procedures where the provenance of the data, i.e. the different integration practices, are not saved. Imagine the following: a researcher is interested in analyzing changes in the housing situation in the Netherlands, prior and after World War II. To answer this question first the researcher needs to spend laborious time just to find out the location of the files he or she is interested in. After

identifying the corresponding files, the data is manually extracted (whether from images or Excel tables in the case of the Dutch census). To answer the research question the data is then transformed (defined, standardized, mapped etc.) and made comparable for that specific question in mind. In other words, the data is interpreted in a way that is difficult to repeat, i.e. according to the view of that specific researcher. Although the outcomes of this work are documented in the scientific literature and disseminated according to best practices, such a question-oriented approach hampers the reproducibility and reusability for other researchers considerably (Denley, 1994, Merry 2006, Boonstra et al. 2006).

For many years, using the Dutch historical censuses has been quite problematic to say the least. In this section we specify the key problems hampering the (re)use of the Dutch historical censuses for comparisons over time and space. We categorize these problems into three main groups. The first problem relates to the fact that the data we are trying to make comparable is only available in *aggregated* form, except the censuses of 1960 and 1971. In fact, the Dutch census mainly provides counts, e.g. "1678" occupied houses in the municipality of Achtkarspelen in 1869 and for most years no micro data was preserved. This lack of micro data necessitates a different approach in order to make the data comparable across censuses. Studying the harmonization of aggregate historical census data across time and space is a terrain not yet explored. The absence of similar harmonization efforts makes this a key challenge to overcome in this research. The second major problem with the census as a source for longitudinal research is related to *changes*. Throughout its existence the censuses have gone through many changes to reflect different

needs, resulting in changing enumeration methods, variables, classification systems and the structure of the tables in which the data were modeled from census year to census year. The third major issue of the census is related to its different *transformations* and the digitization problems introduced during these processes. The problems described with regards to diversity in data formats, structures, context and content of historical censuses calls for a unified system. Data integration and uniform ways of accessing the data is therefore a necessity in order to do any type of longitudinal research. In the following sections we describe why the different problems we have identified often prevented the use of historical census data for longitudinal analysis.

1.3.1 COMPARING AGGREGATE DATA

In contrast to many countries, most of the census data collected in the Netherlands has been preserved on an aggregate level only. The original information collected by the enumerators on sheets were not preserved but aggregated and published in books. The Dutch historical censuses span from 1795 until 1971. From this period we primarily have micro data for the census of 1960 and 1971, made available by the Dutch Statistics (CBS) and DANS. For the years 1830 and 1840 about half of the original census sheets have survived and are available at the municipal archives (Muurlings and Mandemakers 2012). In this study we solely aim to explore and develop methods for comparing historical aggregate (census) data over time. Currently, in the realm of historical census data integration studies there are several successful efforts. These efforts however build on micro data methods but only a few on aggregate data alone. However, comparing micro data over

time entails a different approach compared to aggregate data. The imperative difference between the two is that when using micro data one is able to (re)build classification systems and variables according to one's need. With micro data at our disposal we can go back to the original data and *reclassify* the data in order to create new harmonized variables or classification systems. This could be the case when creating new classification systems for occupational titles, religious denominations, various housing types, different age ranges etc. For example, censuses use different levels of detail to classify the 'age ranges' across the years. With micro date at hand we can reclassify the age ranges as we need to provide maximum comparability over time. This could in practice mean that we use the original data to create new overarching age ranges such as e.g. 15-20, 21-25 and 26-30 which *replaces* the original ranges 15-22, 23-30 for one census year and ranges as 15-18, 19-30 for other census years. The key aspect here is that we can create this new age range by *reclassifying* the micro data, whereas with aggregate data we are bound to interpolations or other statistical estimation methods. The same also applies when dealing with religious denominations or occupations. Throughout the census different levels of detail are used when referring to religions. In the early years of the Dutch historical censuses (i.e. 1830 and 1840) only four religious groups were identified, namely Protestants, Roman Catholics, Israelites and Others. Ten years later the Protestant group is divided into detailed sub denominations such as Anglikaansche Episcopalen or Doopsgezinden. Having micro data we could recreate the subdivisions of the religious denominations of 1830 and 1840 and create a more detailed enumeration for the various religious beliefs to make them comparable with the religious variables of 1850 and beyond.

Building on the examples we described with micro data we now take a look at the main difference compared to having aggregate data as a starting point. In the previous examples we have seen which harmonization options users have when micro data is preserved. However with aggregate data the aforementioned methods do not apply. With aggregate data we cannot simply go back to the original data and reshuffle it into higher or lower level variables. To achieve similar harmonizations with aggregate data we are often forced to create variables which are based on estimations, interpolations and other statistical techniques in order to allow comparability across the different census years. For example, to harmonize the same age ranges with aggregate data we need to apply *interpolations* in order to create harmonized variables for the age ranges 15-20, 21-25 and 26-30 which are based on the original ranges 15-21, 22-26 and 27-32. In the case of religious classes (Knippenberg 1992) which have been splitted into subgroups, i.e. as in the case for the Protestants after 1840, we are forced to *estimate* the subgroups for 1830 and 1840 based on data and ratio from the censuses of 1850 and beyond. The main difference in both scenarios is that we are *creating* overlapping variables across the years based on *statistical estimation*s. Therefore, harmonization of aggregate data introduces more ambiguity and uncertainty compared to micro data practices.

1.3.2 THE CHANGING STRUCTURE OF THE CENSUS ITSELF

Next to the problem of aggregate data, the Dutch historical census itself presents many problems to overcome before being able to use it for longitudinal analysis. In this section we first present

problems dealing with variables and their changing nature. Next we describe problems with regards to how these variables and values are organized in the various classification systems of the census. Consequently, we present the problem of the changing internal structure of the tables, i.e. the way the census was organized.

CHANGING VARIABLES

Changing variables are a key bottleneck preventing researchers to use the historical censuses for longitudinal analysis in an efficient way. Throughout the entire census period the published variables were very much subject to change every ten years. When referring to this problem of changing variables different scenarios can be identified. These represents the different ways in which the census variables tend to behave over time.

A very obvious change scenario, but still difficult to handle, is when the names of the variables are changed from census year to census year. This could be a small variation in the spelling but quite often we encounter variables which completely change to another label. For example a very basic but crucial demographical variable in the census, actual population size, is often referred to differently. The 'actual population' size "juridisch aanwezige bevolking" in Dutch, is referred to as: *Totaal*, *Bewoners*, *Mannen*, *Vrouwen*, *Aanwezig (totaal der feitelijke bevolking)* or *Bevolking die in de gemeente werkelijke woonplaats heeft* etc. As we can see these terms are not very much related lexically. More simple changes are when 'mannen' (males) are referred to as *Mannen*, *Mannelijk* or just *M*. Without expert knowledge or a in depth study of the contents of the census tables, asking a simple question to

determine the actual population size in the Netherlands at a given time does not have a straightforward answer.

Another problem with variables is related to ambiguity. This means that we can find exactly the same label but with a different meaning, sometimes even in one Table but mostly across other years. This is for example the case with the term '*Huizen'* and '*huizen'*, the municipality named Huizen versus the word for houses in Dutch (for clarity and the purpose of this example we have capitalized the municipality). We also find examples where the label "Totaal" has different meanings across other years. In these cases it is the context and expert decisions which determines the actual meaning and helps us to deal with the ambiguity problem.

The foregoing contains mainly examples of variables which use *variations* in labeling. Working with historical census data we find different scenarios where the variables considerably evolve over time. More concretely, the problems users of the census face are: joining two or more variables into one, the splitting of a variable into more detailed variables, the introduction of variables only for specific years or variables which are withdrawn from a census. The latter is the case for the census of 1879 where suddenly the population total is made implicit. In this scenario the variable 'total population' was removed from the tables and needed to be reconstructed by summing up the of total males and females. Other scenarios of variable splitting and joining are one of the most problematic to deal with because of the aforementioned issues related with aggregate data. For example, due to specialization or differentiation occupational categories were often split or merged again for budgetary reasons. Other variables such

as religious denominations and context (i.e. geographical variables) share the same scenario. Religions tend to split or sometimes go together in new branches, making it difficult to trace across time. Looking at the problems with geographical variables such as municipalities we are faced with hundreds of municipalities, their changing boundaries and composition (Boonstra 2006, 2007). Municipalities have been created, merged or split almost constantly throughout history in the Netherlands. In fact, in almost two hundred years there were only six municipalities in the Netherlands which did not experience changing boundaries.

CHANGING CLASSIFICATIONS

The changes in the census and the evolution of the variables are strongly reflected in the different classification systems used in the Dutch historical census. Throughout the censuses various classifications systems have been used to organize all variables and their values in order to put them into meaningful groups. However major changes between the classifications systems used makes it problematic for researchers to efficiently utilize the census for longitudinal studies.

The classification of variables is a necessary step in reducing the information deluge and providing manageable proportions when trying to make sense of a subject matter as a whole. Next to variables with a handful possible values (such as sex or marital status), we often find variables with over hundreds, sometimes thousands of values which need to be grouped in a sensible way in order to study a subject matter. Such variables in the Dutch historical census are for example: occupations, housing types,

religious denominations, geographical context variables etc. The change in classification and level of detail is perhaps the most prominent with occupational variables. Occupations were first introduced as part of the Population census in 1849 and 1859 and were later recorded separately in an occupational census from 1889 onwards. The occupational classification system of 1859 contains 31 classes with a distinction between businesses and industry, containing 379 different occupational titles. The occupational classification of 1899 does not merely contain more classes and occupations, it also provides more detail (introducing new variables such as social/occupational position and subclasses). For 1899 we count 36 classes, 3952 occupational titles, four occupational positions and various values for the different subclasses. To make it more problematic, the occupational census of 1947 contains 29 classes but did not publish any occupational title at all. Dealing with such changes is a major but necessary undertaking, when aiming to analyze occupations across time. The only noteworthy effort in the Netherlands dealing with such issues is that of De Jonge (1966).

In the case of geographical context, the variable municipality has also gone through considerable changes. In order to compare our data across space, the classification of this variable is essential. When municipalities merge, split, emerge or disappear we need a uniform way of accessing them both across time and space. For example when we are interested in the number of 'temporary absent males' in e.g. the municipality of Rotterdam in 1879, we actually want the municipal borders and composition of that period. The city of Rotterdam consisted of Delfshaven, Kralingen and Charlois until 1934. After this year the city was gradually expanded with i.e. Pernis and Hoogvliet in 1934, Ijsselmonde,

Hillergersberg, Overschie and Schiebroek in 1941 and just recently Rozenburg in 2010 (Van der Meer and Boonstra 2006). This example clearly shows the importance and the need for a classification system which keeps track of the composition and borders of municipalities at given times. When dealing with other problematic variables such as 'housing types' or 'religion' the need for pragmatic solutions becomes even more evident. Where in some years housing types are published with just a minimum level of detail such as 'inhabited houses' and 'uninhabited houses', other census years provide a very precise range of housing types.

CHANGING TABLES (STRUCTURES)

Even when the data is standardized and classified according to uniform variables and classification systems, changes in the digitized Excel tables and their varying structure make it difficult for researchers to use and access the data over the years.

The first problem relates to the lack of a connected system which allows us to analyze or access the data as a whole. Over the years the Dutch historical censuses have been converted into Excel and not to a database system. In practice this means that users need to download and search for the data they are interested in by manually opening and closing the different tables. To make it more problematic, there is no clear structure in the way the tables are organized. For some years there are single Excel files containing twelve sheets with a Table for each province and one for the nation as a whole. For other years we find twelve different Excel files with only one Table. This scatteredness of the data results in time consuming data integration and cleaning. For example, when researchers are interested in a simple question such

as ”the total number of inhabited houses throughout 1859-1920”, they need to open 60 different tables and collect the data from 80,032 cells. Even when assuming that the data in the Excel tables are harmonized, researchers still have to extract the data manually. But then, in most cases users do not even know where to start looking and the first basic questions is: in which tables can I find the variables I am interested in? Can we create frequency lists out of the values in order to see what is in there? How are the variables related to one another? Therefore, in practice researchers end up opening *many more* tables than the 60 actually needed as they do not know these answers beforehand.

The second problem relates to structural heterogeneity in the tables themselves. This problem arises with the decision to transcribe the Excel tables in a strict source-oriented manner by preserving the layout of the original census books. Therefore not only are users faced with evolving variables and classifications, they also face changing structures and hierarchies of the layout. The tables are sometimes presented in very simple forms of row and columns with no hierarchy at all and in other years the same data are spread out into more detailed hierarchical tables. These different layouts, which do not follow a pattern, are difficult to align both in context (variables, values, classification systems etc.) as in structures (Table layouts). Therefore, a significant problem researchers are faced with when dealing with the 2,249 Excel tables is how to model the data in a uniform way when moving towards a database system. This modeling is one of the main challenges when moving towards a historical census database, as there is not ‘one’ correct model when comparing aggregate data over time. Although census tables from the *same* year (but of different provinces usually) share the same structure, changes over

the years are evident for almost all tables. Different researchers could therefore have different interpretations on the same data and create diverse models.

1.3.3 TRANSFORMATION PROBLEMS

Other key problems researchers face stem from the various conversions of the census. We distinguish two types of conversion errors. First there are the known errors which were copied from the source material when transcribing the data. Due to the strict source-oriented digitization approach, even the mistakes were digitized. These mistakes could be wrong numbers in the original books such as incorrect totals, missing data for a certain geographical context or under-representation of females in some years. Even handwritten notes that were used to annotate / correct the data in the original books were copied in an inconsistent manner to the Excel tables. Sometimes they were used to change the data, sometimes they were only copied as annotation. And since the process of improvement continued after the initial data entry, it has even become impossible to distinguish which annotations are from the source and which ones are made more recently by the institutions correcting the data.

The second major problem researchers face are the different mistakes introduced during the manual transcription process from the images to the Excel tables. Although great effort was put into representing the source data and structure as closely as possible, this did not go according to plan in several cases. For example, throughout the 2,249 Excel tables we find numerous tables which suddenly use formulas in Excel to calculate the totals instead of

manually transcribed totals. This often does not work out well and users end up with incorrect totals or errors in Excel due to wrong formulas. In other cases we even find non integer values which are mostly the result of incorrect data definitions used in Excel when entering the data. Missing / not included data is also a practical problem when researchers want to use the data. For example, for some years municipalities which are present in the original books are missing. For some reason they were not completely transcribed into Excel. Next to these types of mistakes we find more structural problems hampering the standardization and classification of the data in semi-automated ways. In most of the cases variables and values are organized in clear structures, where each of them have their own column or row. However, in some years the transcribers have created very impractical structures where several variables are combined and displayed in one cell using no consistent way of separating the values. For example, in a single Excel cell we find four different values, e.g. "Amsterdam Kom Bewoonde Huizen Wijk D". This string contains the values for the variables: Municipality, Lower level municipal area, Housing Type and Neighborhood, all in one cell.

The problems we described in this section clearly show the need for greater data cleaning, preparation and integration methods. The difficulty here is to identify mistakes which are not that obvious, i.e. finding missing variables and values, dealing with unstructured annotations, identifying and correcting wrong totals, detecting obvious mistakes such as non-integer values as total numbers for people, children which are transcribed as married, etc. and finding ways to assess the overall data quality.

1.4 GOAL OF THIS RESEARCH: TOWARDS CENSUS DATA HARMONIZATION

In order to use the Dutch Historical censuses for studies over time, to analyze the dataset as a whole and access the data in a uniform manner, users are confronted with the aforementioned problems. These problems need to be addressed and solved before being able to do any type of longitudinal research using historical censuses. Census data 'Harmonization' is the method currently applied by researchers in order to achieve this. Harmonization is built on a set of data integration methods and practices aimed to solve the aggregation, change and transformation problems of historical censuses.

Digitization of historical censuses was the start of moving towards historical census databases for research. Although currently censuses are better preserved and accessible, a pivotal shortcoming thwarting the use of this data for research is related to the lack of harmonization. The problems of harmonization are inherent in the very nature and goal of the censuses, i.e. to track and answer societal needs at *given times* in history. However, while staying true to their decennial obligations of providing relevant data, their use for comparative research became problematic throughout the evolution of the censuses when societal needs and ways of counting changed. The solution of social historians to tackle this ancient problem of census data, and the current standard in the field so far, is the creation of so-called harmonized databases (mostly self-contained practices and workflows). Harmonizing different terminologies, classifications and ontologies is thought to be essential for *any* integrated description of census and historical data in general. As we have presented in the previous

section, different ways of counting and digitization, heterogeneous Table structures, evolving variables, values and classifications systems all need to be harmonized in order to access the data in an unambiguous way over time.

Working in line with projects such as the Integrated Public Use Microdata Series (IPUMS), The North Atlantic Population Project (NAPP), the UK Data Service and others, we aim to provide comparable census data over time and space to stimulate greater use by its community and beyond (social and economic scholars, historians, demographers, epidemiologists etc.). In contrast to earlier harmonization efforts we build our methods on aggregate data and use Semantic Web technologies, more specifically the Resource Description Framework (RDF) as the main modeling technique, making cross-disciplinary contributions. The Semantic Web is "an extension of the current Web, in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee, Hendler and Lassila 2001, p. 1). The Semantic Web is considered the collaborative movement and the set of standards that pursue the realization of this vision. RDF is the basic layer on which the Semantic Web is built. The W3C (World Wide Web Consortium) defines RDF as the standard model for data interchange on the Web and has features that facilitate data merging, specifically supporting the evolution of schemas over time. A promising aspect of RDF is that the definition of the content of a value is not included in the definition of a Table structure which is usually the case with e.g. relational databases. By using RDF the census tables can be represented with diverse RDF graphs that match their diverse structures, without constraints on meeting an overall agreed model.

1.4.1 AN E-HUMANITIES APPROACH

Applying computers in history gained momentum in the 60's and has currently become a common practice. We can consider the field 'computing and history' or 'historical informatics' as one of the first meetings at the cross point of Digital Humanities. As described by Haigh (2014, p. 26), the Digital Humanities is a movement and "a push to apply the tools and methods of computing to the subject matter of the humanities". From the mid 80's 'history and computing' got a strong push with the introduction of personal computers and already at the turn of that decade debates on the application of history and computing started to grow and gain momentum (Boonstra et. al 2004). Nowadays, relational databases have become the standard for representing historical data such as the census. However, this did not happen overnight and was the result of the natural discourse of technology in the ever-evolving field of computational history.

Exploring new and more effective methods and technologies to solve longstanding problems offered by social historical data such as the census is a natural discourse. The interplay between technology and historical research, is one which is more prominent within the field of Digital Humanities, compared to the use of different methodologies applied in the confined domains of the different sciences (i.e. history and computer science). In this research we follow this line of development in the field of Digital Humanities and apply 'new' technologies to solve an old problem, i.e. dealing with changes and comparability over time.

Currently, the application of Semantic Web technology is being advocated in different (historical) fields, varying from structured statistical sources such as census data, to audio visual and textual data. Exploring and applying different types of technology is more a means than an end in research. Finding the best solution to a problem, often means exploring new methods and technologies. The fact that current practices such as relation databases are deeply embedded in the workflows of (social) historians, does not necessarily mean that we have reached an impasse and should not explore new methods which promise to contribute to the same cause. In this research we aim to explore and provide alternative ways of dealing with historical census data harmonization, but also to build on current practices and experiences. All this is done with the goal of contributing to longitudinal analysis and re-use of these sources, which until now lacks a generic and structured approach for aggregate data.

1.4.2 RESEARCH CONTRIBUTION

This research focuses on the theory and practice of data harmonization and aims to deliver generic methods and solution in order to provide greater access to and use of the Dutch historical censuses. Harmonization of such a large scale socio-economic historical dataset over time, using generic methods and principles while building on Semantic Web technologies is a novel approach. Although some efforts have already been made to publish census data using these technologies (see chapter 4), they rarely concern historical data and no generic practices and models have been defined so far. We extend this field of research and introduce a

key concept of historical research into the Semantic Web, namely; change/differences over time. Looking at ontological differences over time and providing generic and transparent ways to align such differences is key in our harmonization approach. By using the historical Dutch censuses as our use case, we aim to extract specific harmonization workflows and methods to lay down the ground rules for other researchers aiming to create similar harmonized historical databases.

We believe that providing generic and transparent ways of bringing together unconnected datasets will contribute to enhanced scholarship. As we will show, current harmonization approaches lean highly towards model / goal-oriented solutions to solve the problems associated with the census. However, the nature of our data calls for a flexible approach which allows different interpretations, transparent harmonizations and preserves the link to the underlying sources at all times (a key requirement in historical research). By applying RDF as the main modeling technique, we want to investigate whether (and in which degree) these requirements can be fulfilled. A harmonization approach where all the decisions are accounted for and the data is easily reusable (i.e. open- practices and data), will contribute to stimulate the use of the census in a responsible way.

1.4.3 RESEARCH QUESTION

Until now, there has been no generalizable research on specific census related harmonization efforts as in the case of Dutch censuses, where we mostly have aggregated data. As we show in this study, this type of data calls for a source-oriented

harmonization approach which mostly lacks in current 'question driven' approaches. Extant literature (whether using traditional methods or Semantic Web technologies) do not provide enough insights into the practice and workflow of (aggregated) census data harmonization. The lack of comprehension into the workflow or harmonization of historical census data is therefore still a bottleneck for many researchers interested in using these data.

This study aims to provide a clear insight into the harmonization process of aggregated historical census data and give concrete recommendations on how to deal with the different types of data found in the census (both methodological as well as practical solutions). Following this thought, the main research question of this study is:

" What is the need for historical census data harmonization from a theoretical and practical perspective and how can Linked Data contribute as a new technology."

Our research question addresses three key aspects of census data harmonization. First it aims to define the gap between current practices applied in various projects and the needs of researchers when dealing with the problems associated with historical census data. We review whether and to which degree the theory and practice of census data harmonization is supported by current methods and technologies. Second it focuses on the practical and methodological aspects of data harmonization and aims to make the process more structured for others. Finally we explore the

suitability of harmonizing historical census data using Linked Data technologies, more specifically RDF and the Semantic Web. We explore the appropriateness of using RDF when the dataset suffers from structural heterogeneity and contains major changes from year to year.

With its specific problems the census data requires a combination of research methods, using both quantitative as well as qualitative approaches in order to get a better understanding of the underlying processes of harmonization. Although we use novel technologies such as RDF from the computer science perspective, the knowledge intensive social historical approach in this research is crucial for providing meaningful harmonizations. In this dissertation we aim to identify the harmonization criteria of historical census data from a theoretical and practical perspective. The first stages consist of a literature study to get a better understanding of the current practices and methods according to both theory and practical cases. We aim to identify existing projects dealing with the same issues and do a synthesis on their main characteristics to identify common practices and workflows. As the main users of our end product, a harmonized database, are mainly socio-economic researchers and historians, their input and practical knowledge when dealing with the Dutch census is collected by way of (semi)structured interviews. The practical side of this project includes (pilot) use cases to give us a better grasp of the data and to try out harmonization methods across a limited number of years. These results help us to not only define and experiment with the workflow of harmonization but more importantly to identify practical data problems with the census by way of an *iterative* and gradual process. The main goal of the pilot

project iterations is to create generic methods and technologies which can be applied and extended to the rest of our datasets.

The scope of this study primarily focuses on the harmonization of historical aggregate censuses and the different approaches applied. The 2,249 Excel tables with Dutch census data are therefore our point of take-off in this research. In this study we do not: deal with the process of digitization of census data, transcribe already digitized images to Excel or apply linguistic approaches on the textual descriptions in the census books. However, as we merely have a sub set of the total census data currently available (which have not yet been digitized or machine processable yet) we develop tools, scripts and flexible harmonization methods to allow the dataset to be expanded in the future.

1.5 THE CEDAR PROJECT

This research was conducted within the context of the CEDAR (Census Data Research) project which was part of the Computational Humanities programme, of the KNAW E-humanities Group in Amsterdam (2011-2016). The Computational Humanities programme consisted of four large projects, selected on the basis of international peer review. These interdisciplinary projects involve cooperation between different institutes and universities. The CEDAR project builds on two Ph.D. projects, running in parallel, with the goal of harmonizing and interlinking the data in the Semantic Web. The team consists of an inter-disciplinary group of researchers such as computer scientists (VU), archivists and care takers of the census since the start of the digitization efforts at DANS (Data Archiving

Networked Services), social historians (Erasmus University, Radboud University and International Institute of Social History in Amsterdam) and historical informatics specialists (Radboud University). In this research project we followed a cross-disciplinary and coupled Ph.D. approach. Our aim was to combine and further progress current computational history (historical informatics) efforts with new research fronts in computer and information sciences (i.e. the Semantic Web).

In CEDAR we use the Dutch historical censuses (1795-1971) as a starting point to create generic and integrated methods for comparing structured historical sources such as census data over time, and contribute to scholarly practices with regards to traditional historical techniques. Our contributions are directed at the domains of (social) history, historical information sciences, computational sciences and their combined domain known as digital humanities. By creating a structured, accountable and transparent harmonization model for historical censuses the CEDAR project aims at providing a bedrock for other researchers within the humanities who face similar problems. Comparability across time will push the boundaries of these sciences and we aim to inspire others in the value of interconnecting historical sources over time in structured and transparent ways, harmonizing unlinked datasets, bringing datasets together to embed external sources and make census data inter-linkable with other hubs of historical socio-economic and demographic data and beyond.

The focus of CEDAR is on knowledge representation and on providing generic ways of integrating dispersed datasets (specifically the historical censuses) and on the other hand on how to achieve this by using Semantic Web technologies. Harmonizing

a large social economic historical dataset such as the Dutch census data in RDF is a first. The lessons learned from this research can benefit others in their efforts to do the same. In this project we pave the way for other digital humanities projects in harmonizing their data and more efficiently publish and (re)use them in the Semantic Web.

1.6 CONTENT OF THIS STUDY

This dissertation builds on three main parts, some consisting of several chapters. We first give a general introduction of each part and consequently provide a more detailed overview of the different sections and chapters. Here we also present the sub-questions we cover in the different sections of this study.

The first part consists of two chapters and describes historical censuses, their potentials and limitations, the different approaches taken by other census harmonization projects and introduces our view and definition of census harmonization. We start with describing the censuses, their importance throughout history and the potential they harness for research purposes (chapter 2). To understand the scientific value of the censuses we go into the problems and challenges that the use of this data entails. Here we will already discuss the direction into which possible solutions may be found. In chapter 3 we take a closer look at the theory of census data harmonization. When moving towards a harmonized database from unconnected files with different degrees of detail the source and goal-oriented paradigm are prominent topics of discussion in the field of history and information sciences / computational history. We look at current efforts in the

traditional approaches as well as attempts in RDF and build on these experiences to identify a set of harmonization practices before defining a generic and transparent workflow. We close this chapter by providing a clear definition of the still illusive concept of harmonization.

Our concept of harmonization is formed in line with current approaches and explores the uses of Semantic Web technologies in historical research. In the second part of this study, chapter 4 and 5, we take a closer look at the practices and challenges in historical research in relation with Semantic Web technologies. These chapters focus on the Semantic Web through a survey and the use of historical data currently applied in this field.

In the third part, chapter 6 and 7, we look into the practice of harmonizing the Dutch historical censuses in RDF, our proposed three-tier model, source-oriented methods and iterative and accountable workflow. We present specific harmonization models and practices of (historical) census data harmonization through a use case and describe a structured approach. Such an approach is desperately needed when dealing with harmonization of such complex data, especially when using RDF which is not meant as an intuitive language for humans to process. We propose a source-oriented workflow and present the outcomes in the form of harmonized tables, visualizations, cross collaboration etc.

In chapter 8 we present the main results of our study. We present a critical review and identify open challenges as a result of our harmonization efforts. Finally we conclude with our findings, lessons learned and contributions in advancing the use and accessibility of historical census data.

1.6.1 PART 1: HISTORICAL CENSUSES AND DATA PROBLEMS: ITS CHALLENGES AND POTENTIALS

Censuses are one of the most difficult data to use for longitudinal studies. To understand why the census got this reputation we look at the census from its early beginnings to the many efforts put into improving the usability of the data. We first describe the role of censuses from ancient times onwards (2.1), whether having negative connotations or used as a valuable tool in answering the changing needs of nations to improve the quality of life. We next focus on our specific dataset and problems, i.e. the Dutch historical censuses (2.2). We describe the history of the traditional door to door censuses held for almost two centuries, the data produced in the form of different censuses and their characteristics. In the following part (2.3) we take a closer look at the source-oriented digitization process of the Dutch historical censuses and the limitations of the produced data. We describe the different sources and how the data have been transformed over the years. Here we show, how these efforts have resulted in thousands of heterogeneous, unconnected Excel tables, structures, classifications and variables. In the following part (2.4) we introduce the various problems and challenges associated with the harmonization of the Dutch Historical censuses and give specific examples which hamper the use of this data by researchers.

In chapter 3 we connect the specific challenges facing the Dutch census data with respect to the theory and practice of census data harmonization. We first look at similar harmonization efforts (3.1) and categorize them according to three major characteristics. This part is followed by the old discussion of the 'source and goal-oriented paradigm' (3.2). We describe the scientific discourse of

historians and the decisions taken when moving from source data to historical databases and the need for more source-oriented harmonization approaches. In section 3.3 we describe the theory of harmonization, and give a (census) specific definition, allowing us to get a clear comprehension of its concept before introducing the Semantic Web and how we aim to create a structured and accountable workflow in the following sections.

SUB-QUESTION ADDRESSED IN PART 1:

- Which purposes did the censuses serve in the past?
- What is the history of the Dutch censuses?
- Which characteristics hamper its current use?
- What are the characteristics of current harmonization practices? i.e. can we identify common practices and methods ?
- To which extent do current harmonization approaches support / and or provide generic methods?
- What is the relation between the key aspects of historical research, i.e. change over time, and harmonization?
- What is the difference between source-oriented and goal-oriented methods?
- Why is a flexible approach necessary when harmonizing aggregate data?
- What are the key harmonization criteria according to the principals of historical research when dealing with transformations and analysis of aggregate historical statistical data?

- How can we define harmonization of historical aggregate census data?

1.6.2 PART 2: HISTORICAL RESEARCH IN THE SEMANTIC WEB

Part 2 is divided into two chapters (4 and 5). To harmonize the Dutch historical censuses we apply a specific knowledge representation model from the Semantic Web: the Resource Description Framework (RDF). The RDF framework promises to provide better access to and use of the historical censuses using Linked Data principles. Semantic Web technologies are currently advocated and applied in a number of situations, environments, applications of historical computing and historical information science. In chapter 4 we present the concept of historical information and the Semantic Web. This chapter first focuses on the scholarly practices of historical research (4.2). Next we introduce the Semantic Web and its uprising the field of historical research in section 4.3. In order to understand the nature and challenges of historical data we present the historical life cycle and classification of data in 4.4. We end this chapter with an overview of historical data problems, for which the Semantic Web provides possible solutions in chapter 5. Here we provide a survey covering current Semantic Web technology developments, as well as typical scenarios in which these technologies are and could be currently applied. Combining the various contributions made so far in a single view enables us to grasp the diversity on the borders between historical computing and Semantic Web research. Although in our research we focus on structured statistical

historical data, we want to provide insights into the greater applicability of Semantic Web technologies in historical research. We present the different applications of historical knowledge modeling in chapter 5.1 and different integration efforts of historical sources in 5.2. The following sections (5.3 and 5.4) present open challenges in solving historical research problems. In 5.5 we appeal for flexible data models when moving toward historical databases and wind up with lessons learned from this survey in 5.6.

SUB-QUESTION ADDRESSED IN PART 2:

- What is the information life cycle in the field of historical research?
- What is the Semantic Web?
- How is historical data classified?
- What are the traditional challenges in historical research?
- What is the current landscape of historical research and semantics?
- What are current data integration problems in historical research?
- Which contributions have been made with regards to historical knowledge modeling?
- What is the role of ontologies and of classifications systems in the harmonization of historical Census data?
- How can Semantic Web technologies assist in the data linkage and integration process?
- How does RDF enrich current methods and models when going from disparate files to database?

1.6.3 PART 3: THE PRACTICE OF HARMONIZING HISTORICAL CENSUS DATA: A FLEXIBLE AND ACCOUNTABLE APPROACH IN RDF

The fourth section (chapter 6 and 7) combines the need for harmonization of historical censuses and the solutions in the Semantic Web. After introducing the concept of the Semantic Web we focus specifically on its application and uses for harmonization of historical census data across time. In the following part (6.1) we look at the specific use of applying Semantic Web technologies for historical census data harmonization. In section 6.2 we present a 'three tier' harmonization model, which will serve as the basis of our approach in RDF. We next describe (6.3) our straightforward process of converting thousands of heterogeneous Excel tables, as one to one (source-oriented) copies into a RDF database. In section 6.4 we show the preliminary uses of having the raw census data in RDF.

In chapter 7 we explain how we have implemented a source-oriented, structured harmonization approach with the Dutch aggregated historical census data, using Semantic Web technologies. We propose an iterative and accountable harmonization workflow in RDF (Resource Description Framework) which makes different interpretations possible without losing track of the original aggregated data. The source-oriented approach is the preferred and preeminent method in historical research. Most harmonization efforts, however, currently lean towards goal-oriented solutions. We start with the need for source-oriented harmonization approaches when dealing with aggregate historical (census) data in 7.1. Next we go into the

details of the different stages of our harmonization workflow (7.2), explain each step of our suggested approach (i.e the ground rules) and how we have used this workflow to gradually build harmonized tables in the context of a pilot project. We describe our pilot project and use this generic workflow, to (further) develop, test and explore our methods and data, for the so-called Local Division tables of the Dutch population censuses. Using RDF as the publication and harmonization model we aim to provide full transparency (7.3). We provide accountability (provenance) on two key levels, i.e. the harmonizations applied to the data and the link to the original underlying sources. In section 7.4 we show some statistics with regards to the harmonizations we have made. We conclude this chapter in 7.5 by showing the different contributions and some practical uses which have been already made with a source-oriented harmonization approach in RDF.

SUB-QUESTION ADDRESSED IN PART 3:

- Which harmonization model is best suited for census harmonization in RDF?
- How to represent historical census data in RDF format?
- Which processes / workflow can be identified for the harmonization of historical census data?
- What are the key criteria when dealing with aggregate historical census harmonization (i.e. the practice)?
- How does RDF differ in contrast to traditional harmonization efforts when moving towards historical databases?

- To which extent are historical ontologies and variables available for reuse from the Semantic Web?
- Why is accountability / a transparent approach important when dealing with the harmonization of aggregate historical data?
- Which workflow and activities can be identified when dealing with historical Census harmonization?
- Is RDF meant to assist or lead the harmonization process?
- Which other fields of research and datasets can benefit from a generic harmonization approach?
- What kind of classification systems does the Dutch historical census necessitate?
- How to connect internal classification systems of censuses?
- How do semantic technologies play a novel role in solving and implementing harmonization for historical Censuses?

The final chapter (8) ends this study with a critical review and evaluation of our findings and proposed methods, a description of the open challenges and a discussion of the results and the wider impact of our structured source-oriented harmonization system.

1.7 SHARED WORK AND PUBLICATION OVERVIEW PER SECTION

The dissertation is composed of text not published before, indicated with 'original text' in the Table overview and of text from four articles of which two have been published in key

journals in the field of historical research and the other two published in a well-known Semantic Web conference and journal. In this section we present the main publications which are used as parts of some chapters in this dissertation. We also state which parts of this dissertation is based on collaborative work.

SHARED AND REUSED WORK

THE AGGREGATE DUTCH HISTORICAL CENSUSES: HARMONIZATION AND RDF (published in Historical Methods)

This article describes the Dutch historical census, its history, the challenges faced when using these data for comparisons over time and the possible solutions. We look at the possibilities of doing this in RDF and create a specific model to do so. Ashkpour was the first author and main contributor of this article. Everything except the technical descriptions are original text contributed by Ashkpour.

Ashkpour, A., Meroño-Peñuela, A., Mandemakers, K. The Aggregate Dutch Historical Censuses: Harmonization and RDF. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 48(4), pp. 230-245. (2015).

SOURCE-ORIENTED HARMONIZATION OF HISTORICAL AGGREGATE CENSUS DATA: A FLEXIBLE AND ACCOUNTABLE APPROACH IN RDF (published in Historical Social Research)

This article combines the challenges presented in the Historical Methods article with the opportunities which arise with Semantic Web technologies as described in the Semantic Web journal publication. In this article we adhere to the preferred method in the field of historical research and present a much needed approach which we refer to as source-oriented harmonization. This article presents a structured source-oriented harmonization workflow which can be used to iteratively explore and harmonize data, as complex of historical census data, in an accountable way. Ashkpour is the main author and contributor of this article.

Ashkpour, A., Mandemakers, K., Boonstra, O. Source Oriented Harmonization of Aggregate Historical Census Data: a flexible and accountable approach in RDF. Historical Social Research / Historische Sozialforschung, 41(4), pp. 291-321. (2016).

SEMANTIC TECHNOLOGIES FOR HISTORICAL RESEARCH: A SURVEY. (published in Semantic Web Journal)

This article gives insights in the use of semantic technologies in the field of historical research. In this article Ashkpour was responsible for the historical contents, describing the work processes of historians, the information life cycle, the role of classifications and its application on different type of data.

Ashkpour also was responsible for co-creating the survey itself and providing descriptions of other projects.

Meroño-Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., van Harmelen, F. Semantic Technologies for Historical Research: A Survey. Semantic Web – Interoperability, Usability, Applicability, 6(6), pp. 539– 564. IOS Press. (2015).

FROM FLAT LISTS TO TAXONOMIES: BOTTOM-UP CONCEPT SCHEME GENERATION IN LINKED STATISTICAL DATA (published in SemStats 2014 proceedings)

This article looks at the possibilities to assist researchers in creating classification systems for variables which have many possible values. The contributions of Ashkpour were related to the creation and curation of expert classification systems. These expert classification systems were used as the 'golden standard' in this study and used to compare the results of the automated classification suggestions. More specifically, a detailed housing classification was built by Ashkpour using data from several census years.

Meroño-Peñuela, A., Ashkpour, A., Guéret, C." Proceedings of the 2nd International Workshop on Semantic Statistics (SemStats 2014), ISWC 2014, Riva del Garda, Italy (2014).

OVERVIEW OF TEXT USED PER SECTION AND CHAPTER

This Table presents an overview of the different chapters per part and which texts and publications are used.

Part	Chapter	Source
Introduction	1	Original Text
1	2	Original Text
		HM
	3	Original Text
		HSR
2	4	Original Text
		SWJ
	5	Original Text
		SWJ
3	6	HM
		ISWC
	7	HSR
		Original Text
Conclusion	8	Original Text

Table 1.1 - Overview of sections, chapters and text used.

HM: Journal of Historical Methods
HSR: Journal of Historical Social Research
SWJ: Semantic Web Journal
ISWC: International Semantic Web Conference publication

Part I

HISTORICAL CENSUSES

2. HISTORICAL CENSUS DATA

Before going into the details and peculiarities of the Dutch historical censuses, we first take a step back and look at several key historical censuses and which purposes these censuses served. In section 2.1 we start with describing the role of censuses throughout ancient times (3800 B.C.E) until the end of 18th century. After this general introduction into the different uses of census data throughout history we focus on our specific dataset and problems, i.e. the Dutch historical censuses (2.2). We describe the traditional door to door censuses held for almost two centuries and the data they produced with their different characteristics. In the following part (2.3) we take a closer look at the digitization process of the Dutch historical censuses which started in 1997 and is currently still ongoing. The produced digitized dataset, namely: thousands of heterogeneous, unconnected Excel tables with all its limitations and possibilities, are our point of take-off in this research. In the following part (2.4) we introduce the various problems and challenges associated with the harmonization of the Dutch Historical censuses and give specific examples which hamper the use of this rich data source by researchers. We also already present possible solutions and requirements when harmonizing aggregate historical censuses.

This chapter is based on the following work. (1) Ashkpour, A., Meroño-Peñuela, A., Mandemakers, K. The Aggregate Dutch Historical Censuses: Harmonization and RDF. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 48(4), pp. 230-245. (2015). This article contributes

to this chapter by describing the history of the Dutch censuses (2.2) and identifying the main bottlenecks and needs for harmonization. This section is a strongly extended version of the content presented in the introduction. For this chapter only the first parts or this article are used. (2) Original work specifically written to enrich this chapter with extra background information on the historical aspect of the censuses. Original work was written for section 2.1 describing ancient censuses throughout history and extra background information was added to all the different subsections of this chapter.

2.1 CENSUSES THROUGHOUT HISTORY

In this section we show the role and importance of census data throughout history. We describe how the negative *perception* of ancient intrusive census taking methods echoed over the years and how a shift in enumeration goals gradually took place over time. Throughout history the census developed from a tool which was mainly used for governing purposes (taxation and military) into a tool for statistical observation.

The word census finds its origins in the Latin word "censere", meaning to *assess* or to *value*/*tax*. Whereas modern censuses aim to include all persons in the nation, ancient censuses were mostly limited to counts of taxpayers (merchants, farmers, landlords etc.), heads of households, males of military age etc. (Kaplan 1980). Moreover, women and children were rarely included. It is not surprisingly then that the word census historically had a negative, somewhat forced connotation to it. It was associated with war or taxes, both things which people wanted to avoid.

Counting populations and the use of statistics for different purposes has been a recurrent theme in history. The first known census was taken in the Babylonian empire around 3800 B.C.E. According to records these censuses were taken every six or seven years (Grajalez et al. 2013). The Babylonian censuses counted among others the number of people but also property such as the number of livestock, quantities of honey, butter, wool and vegetables to measure its wealth and resources for the kingdom. This information was used to e.g. divide food and resources among the population. An example of this ancient census has still survived and is currently displayed at the British Museum. Around

500 B.C.E one of the earliest *documented* censuses was carried out in the Persian Empire (Kuhrt 1995) for issuing land grants and taxation purposes. The earliest still *preserved* census, is that of the Han Dynasty. Taken in China in 2 CE, it is believed to be a very accurate enumeration of its population (Kaplan and van Valey 1980) and included information on the, households as well as on the population of the country as a whole (Durand 1960). The second oldest still preserved ancient census also comes from the Han Dynasty taken around 140 C.E. This census records the enormous decline in the population size during that time. Due to mass migrations into what is today southern China, the population size diminished from around 59 million to 48 million people.

In the Bible ancient censuses are mentioned several times. The first one is at the time of the Exodus around 1490 B.C. The Book of Numbers starts with a divine order leading Moses to count all men from 20 years upwards. The Lord spoke to Moses in these words (Kaplan and van Valey 1980):

"Number the whole community of Israel by families in the father's line, recording the name of every male person aged twenty years and upward fit for military service".

Following this, Moses and Aaron summoned the whole community on the first day of the second month, and they registered their descent (The New English Bible, Numbers 1:1-10). This enumeration revealed a total of 603.550 men of the age 20 and above.

Around 500 hundred years later the Bible mentions a second census. On the order of King David of Israel a census was taken around 1000 B.C, against God's will. The King instructed Joab and the officers of the army to:

"Go, number Israel from Beersheba even to Dan; and bring the number of them to me, that I may know it" (King James Bible, 2 Samuel 24:1-9).

Under protest but to no avail, they covered the whole country and arrived back at Jerusalem after nine months and twenty days. Joab reported to the king the total number of people: the number of able-bodied men, capable of bearing arms, was eight hundred thousand in Israel and five hundred thousand in Judiah (Kaplan and van Valey 1980). By taking this census King David (knowingly) went against God's will and a divine punishment was the consequence where around 70,000 people died of the plague (NKJV[1] 2008). After this, taking a census also became known as the 'sin of David' in western Christian nations (although he is also known for other sins, e.g. laying with Batsheba, wife of Uriah, and sending him to a certain death in the frontlines of war).

Perhaps one of the most famous mentioning of a census is in the Gospel of Luke, where Joseph had to travel to Bethlehem in order to register with Mary and to be enumerated in a census issued by the Roman Emperor Augustus. A Decree was issued by the Emperor for a registration to be made throughout the Roman world. For this purpose everyone made his way to his own town

[1] NKJV: New Kings James Version

(Kaplan and van Valey 1980; Taylor 1933; The New English Bible Luke 2:1-5).

"And everyone went to their own town to register. So Joseph also went up from the town of Nazareth in Galilee to Judea, to Bethlehem the town of David, because he belonged to the house and line of David. He went there to register with Mary, who was pledged to be married to him and was expecting a child." (Luke 2:3-5)

The Roman censuses were one of the most developed censuses in the Ancient world. Starting around 550 B.C. with enumerating only districts near to Rome, these censuses were gradually extended to cover the entire Roman Empire by 5 B.C. The Roman censuses are recorded to have been held once every five years (Scheidel 2009), counting the *population* and their *property* for taxation and military purposes.

In ancient times the Greek believed in the concept of an *optimum* population size. According to Plato the ideal city should have a population of "5040" citizens, however, excluding women, children and slaves from this count (Sharma 2007). Although even not in the power of Plato to realize this number, some necessary 'encouragements' were made to control its population growth according to this ideal of an optimum population size. These encouragements could be rewards to make sure that the land was divided between 5040 citizens via an inheritance system, pressures to keep the offspring level at the desired number (e.g. advice from the elders), social stigma or even sending of the excess numbers to a colony (Daugherty and Kammeyer (1995).

In medieval Europe, a notable historical census is the one that resulted in the famous Domesday Book. This enumeration was ordered by William the Conqueror of England in 1086, to assess

and tax his newly conquered land (Quinlan 1990). This census was more focused on land than on enumerating the people and was one of the first censuses in the western world after the fall of the Roman Empire.

Looking past the ancient censuses and towards more modern ones we see a shift in the use and goals of the census, more aiming to serve societal information needs rather than war or tax purposes or the often senseless agendas of rulers (e.g. Plato, William the Conqueror). Although not a census, we see practical uses of statistical information to serve societal needs, i.e. using statistics to manage the recurrent bubonic plagues threatening the citizens of London. For example in England, following the London plague of 1603, the importance of gathering information to answer pressing questions was acknowledged by reintroducing the weekly Bills of Mortality (Graunt 1977). From 1629 onwards the cause of death was added to the list, recording all deaths in the city of London. In case of the United States, after the independence war, the nation was formed out of separate states. In order to decide how many representatives the states should send to the U.S congress, a count of the population was needed regularly. This enumeration of the population is even mandated in the constitution[2], Article I, Section 2: "*Representatives and direct taxes shall be apportioned among the several States […] according to their respective numbers*". In this example the introduction of a census played an important role in the constitution of a newly formed nation. The first U.S census was taken in 1790.

Around 1800 the first 'modern' historical censuses were introduced in different *European* countries. The censuses in

[2] http://www.archives.gov/exhibits/charters/constitution_transcript.html

Sweden (1749) and Norway can be seen as forerunners to this development (1760). The first integral enumeration of the entire population in the Netherlands took place in 1795, five years after the first U.S Census. In the following years collecting statistics about ones population gradually spread out over different countries and introduced other additional and essential topics such as in depth demographic information, household compositions, occupation and labor force participation, social equality, housing needs etc.

Looking back at important contributions in the study of censuses and methodologies, one of the most notable ones was that of Adolphe Quetelet. In a time where statistics and social sciences were not seen as a match, Quetelet was the first to apply these methods on such data. Around the time of the industrial revolution (Bethlehem 2009) rapid changes were following each other in different aspects of society and more information about the population was required. Quetelet pioneered in using statistics to answer pressing societal questions, an approach which was primarily applied in astronomy at that time. He argued that statistical models and probability are very useful tools to describe various social, economic and biological phenomena. One of his most influential works of the 19th century is "A Treatise on Man and the Development of his Faculties" (1842). Quetelet applied his knowledge of statistics and mathematics to identify the 'average man' by the mean of a set of variables. By doing so he could compare the features of individuals against 'the average man' and for example identify outliers / irregularities. Looking at crime rates Quetelet was amazed that we now could enumerate in advance, i.e. we could predict:

"*how many individuals will stain their hands in the blood of their fellows, how many will be forgers, how many will be poisoners, almost as we can enumerate in advance the births and deaths that should occur"* (Paulos 1998, p. 169).

Quetelet also pioneered on using the same methods in cross-sectional studies of human growth. Perhaps one of his most famous measures, something still used to this date, was the creation of the Body Mass Index (BMI[3]). Adolphe Quetelet also actively advocated the value and importance of having comparable censuses among the different countries. He was one of the founding fathers of applying statistics in social sciences and great a proponent of more cooperation among statisticians and international bodies. He was responsible for the very first International Statistical Congress in Brussels in 1835 and contributed to the establishing of the CCS (Centrale Commissie voor de Statistiek) in the Netherlands in the same year. Quetelet was also a contributing founder of one of the first historical censuses in the Netherlands (1829) which was based on a Royal Decree.

[3] https://nl.wikipedia.org/wiki/Queteletindex

2.2 THE DUTCH HISTORICAL CENSUSES

2.2.1 INTRODUCTION

Censuses are taken regularly by governments to gain a better understanding of populations and their different characteristics such as size, age-structure, household compositions, occupations and other socio-demographic aspects. The Dutch government collected census information not only to get a view of the state of the nation, but since 1850 also to facilitate the construction and updating of the population registers by the municipal authorities (Den Dulk and Van Maarseveen 1999). Although sometimes lagging behind social reality, historical censuses contain specific information about a nations *population characteristics* and *needs* at a given *time* in history. These needs are for example information about the changing occupational structure and labor force, the housing needs (shortage) after the second world war, the ageing of populations and related issue. For the period before the 20th century, the census is one of the few large-scale historical statistical data sources on population characteristics which are not strongly distorted, providing comprehensive geographical coverage (Ruggles and Menard 1995).

2.2.2 BACKGROUND

The first integral enumeration in the Netherlands started in 1795 under the French influence during the Batavian Republic. Its purpose was to collect quantitative information in order to create a new system of electoral constituencies. It took over 35 years before the next general Population Census was organized. In 1829

the first general census was taken by the Dutch government, based on the Royal Decree of 1828 signed by Willem I der Nederlanden[4] on September 29. The Royal Decree was later replaced by the 'Volkstellingenwet' of 1879. This law stated that starting from 1879, the census was to be taken every ten years (meaning the years ending with the number '9'). The main reason for replacing the Royal Decree with the 'Volkstellingenwet 1879' was that only through legislation, penalties were found to be valid (van Maarseveen 2002). This practically meant that upon refusal of participation a maximum of two weeks in custody or a fine of 500 guilders could be given. This law lasted for almost a century and was enriched by the 'Wet of 1918[5]'. The main goal of this was to improve the demographical comparability of the outcomes with other nations. Another step to align the Dutch Census with international efforts was to change the enumeration year to all years ending with a '0'. Consequently, from 1920 onwards the Dutch census was taken every ten years with an exception due to the Second World War.

After the war the need for housing and population statistics was dire. Even more since the most recent data were 17 years old, due to the cancellation of the census of 1940. Therefore, the census of 1950 was already taken in 1947. The census of 1960 continued in its normal form. After this period the original law of 1879 was entirely revised and replaced by the '1970 Law[6]', where more attention was given to privacy matters and other methodological aspects with regards to the collection of the data (van Maarseveen

[4] https://nl.wikipedia.org/wiki/Willem_I_der_Nederlanden
[5] Wet 1918. Staatsblad 1918, nr. 270.
[6] Wet 1970. Wet houdende regelingen betreffende algemene volkstellingen (Volkstellingenwet) Staatblad 1970, nr. 323

2002). Although privacy was a high priority for the CBS a general privacy legislation was lacking and confidentiality provisions were not part of the 'Volkstellingenwet 1879'. The new '1970 law' however could not withstand the increasing social and political pressure. Due to more awareness and protest with regard to privacy matters, but also political and budgetary aspects the last 'traditional' door-to-door census was held in 1971. Although a high non-response was feared, only 2.3% of the population refused to be counted in one way or the other. As an interesting side note, one of the main objectives of the census was to check the population total according to the 'de jure' concept (see 3.2.1). This concept looks at persons who 'ought' to be registered in the population register. One of the reasons for the many protests (including violent ones) against the census in 1971 was related to this *checking* of the population. Protesters were concerned that the guarantees for the protection of privacy were unsatisfactory (Katus, 1984).

The 1971 census marks the end of the traditional census in the Netherlands which in total covered 17 census years for almost two centuries (Den Dulk and Van Maarseveen 1999). After the final (door to door) Census of 1971 the 'Volkstellingen Laws' were ultimately withdrawn due to fears of a low participation in the next census of 1981 (van Maarseveen and Doorn 2001) of which pilot studies suggested a 26% non-response rate. In 1980 the Minister of Economic Affairs suggested to postpone the census due to these problems and noted:

"*A social-pshychological climate has arisen that is adverse to taking a successful census*" (HTK, 1980b[7]).

On 29 September 1981 of October his proposal was passed by the First chamber and the end of the historical (door to door) censuses became reality. However, this did not exempt the Dutch government in its obligation to meet European regulations and to collect this type of information about its population. Permission was given by the EU to collect the required data by means of a compensation programme. This programme allowed combining data from integral registrars (Dynamic register of the population per municipality) and large sample-surveys for all future censuses. The 'Census Law' was officially revoked in 1991.

Unfortunately, because of the existence of the population registrars from 1850 onwards, the original census forms (1850-1947) were *not preserved*. However, from the earlier censuses (1829 and 1839) about 50% of the nominal manuscripts are still kept in local archives (Muurling and Mandemakers 2012). For the last two census years, 1960 and 1971, the micro-results have been preserved on tape (Van Maarseveen and Doorn 2001). For the period 1850-1947 the results of the census are only preserved at the *aggregated* level and published as tabular data in books. Although these books have been one of the most consulted sources of statistics in the Netherlands and have become a valuable source of information for researchers, the use and accessibility of these books is quite problematic and therefore limited.

[7] HTK 1980b. Handelingen van de Tweede Kamer 1979–1980, 15800, nr. 91 (Rijksbegroting 1980, 6
Hoofdstuk XIII)

2.2.3 CENSUS TYPES

When referring to the published results of the Dutch historical census data over the years, we have to distinguish three main types of aggregate census data: population, occupation and housing data. The history of the Dutch censuses started with a general population census. The occupation and housing census were at this stage not distinctively present. The *type* of information which had to be collected in each census was defined by way of an 'Order in Council' (Algemene Maatregel van Bestuur).

Published tabulations on the *population* span the entire range of our historical census dataset (1795-1971). The population census contains information on the key characteristics of a nation such as gender composition, age, marital status, household situation, nationality, religious divisions etc. The *occupational* census tables were published for the censuses of 1849, 1859 and from 1889 onward. The introduction of the occupational census was inspired by the emerging modern industrialized society which created the need for information about the labor force. *Occupation* is one of the most unwieldy variables in censuses compared to other variables. In fact, the reason why we have no occupational census for the years 1869 and 1879 is because many doubted and argued on the actual use and common sense of this information after the first two occupational censuses were held (Verslag[8] 1888). Historical research on occupations is heavily hampered by misunderstandings with regards to occupational terminology across time and space (Van Leeuwen, Maas and Miles 2004).

[8] Verslag 1888. Verslag van de Algemeene Vergadering der Vereeniging voor de Statistiek in Nederland (1888). In: Bijdragen van het Statistisch Instituut, jrg. 4, 1888, 199–266

Despite such problems, occupation remains a tremendously useful variable that provides invaluable information. For example, age, sex and race are key characteristics intrinsic to an individual. Family relationship describes a person's position within the fundamental social unit. But, only occupation gives an individual's relationship to society in such a succinct and powerful way (Sobek and Dillon 1995). The quality of the early occupational censuses however was quite limited. In previous censuses such as 1849 and 1859, very simple classification systems where used to group the different occupations. In these systems, all people with the same profession, regardless of the business where they worked, were counted together.

The year 1889 was a defining moment in the history of the occupational census. In 1888, the economist and later on first Director of the CBS Anthonie Beaujon (Van der Bie 2014) successfully pleaded for a significant quality improvement of the occupational censuses. In his plea he noted the importance to improve such data, since future social labor laws with regards to (employee) insurance, female and child labor etc. had to be based on solid social knowledge (van Maarseveen 2002). This led for example to the introduction of ages and positions in the occupational census with regards to child labor concerns.

Up to and including 1930, information on the *housing* situation of the Netherlands was collected via the population census. The censuses before this year contain some information with regard to the housing situation and were referred to as 'housing statistics' in the population census. The alternative, i.e. to have a separate housing census with the 1920 population census was declined. The commission indicated that it was preferred to continue with

the housing statistic as part of the population census. The reason for this was that the required type of information was supposed to give insight into the *housing situation* in connection with the persons living in these houses. A distinct housing census would focus more on the actual housing market and housing reserves. The subject of study would therefore be the house itself (e.g. location, number of rooms, type of house etc.). According to the commission a combination of these two approaches was therefore not feasible (van Maarseveen 2002). The first official housing census was introduced in 1947 and was held together with the population census. In 1956, the housing shortage and need for data about the housing market called for a new housing census and this one was conducted separately, independent of the population census. The last official housing census was held in 1971, again together with population census. In Table 2.1 we provide a complete overview of the different censuses taken in the Netherlands throughout the period 1795-1971.

Year	Census
1795	First general enumeration in the Dutch Republic
1829	First General Population Census (by Royal Decree)
1840	Second General Population Census
1849	Third General Population Census (incl. Occupational Census)
1859	Fourth General Population Census (incl. Occupational Census)
1869	Fifth General Population Census
1879	Sixth General Population Census
1889	Seventh General Population Census (incl. Occupational Census)
1899	Eighth General Population Census (incl. Occupational Census and Housing Statistics)
1909	Ninth General Population Census (incl. Occupational Census and Housing Statistics)
1920	Tenth General Population Census (incl. Occupational Census)
1930	Eleventh General Population Census (incl. Occupational Census and Housing Statistics)
1947	Twelfth General Population Census (incl. Occupational Census)
1947	First general Housing Census
1956	Second General Housing Census
1960	Thirteenth General Population Census (incl. Occupational Census)
1971	Fourteenth General Population Census (incl. Occupation Census)
1971	Third General Housing Census

Table 2.1 - Overview of the Dutch Historical Censuses (van Maarseveen 2002).

2.2.4 OBJECTIVES OF THE CENSUS

The objective of the Dutch historical censuses was always twofold (van Maarseveen 2008). The first objective focuses on the *supply of statistical information*. Its aim is to gain insights into the demographics of the population and to assess the size and distinct categories on a certain date in time. From 1850 onwards, the second objective of the Dutch census is *administrative* and focuses on the construction and updating of the population registrars by the municipal authorities.

Supply of statistical information: concepts of counting the population

The objective of censuses taken before 1829 (i.e. the very first integral enumeration of the Netherlands in 1795), were primarily related to establishing electoral districts and taxation purposes. *After* the census of 1795 (in line with the first general Population Census of 1829), the focus shifted towards getting insights on the population demographics and to determine the actual population size. This however proves not to be straightforward as the population size, a key demographic variable, could be determined using different definitions and concepts. The big practical problem is who are actually counted? What to do with individuals who are *temporarily* present during the enumeration but live elsewhere? How to count temporarily absent individuals who are not in the location where they were originally registered? What to do with people who live at a certain residence but are not registered in the population registrar? These questions present different ways in which the *population size* can be assessed (van Maarseveen 2002). The Dutch historical censuses have used the

'de jure' and 'de facto' concepts when assessing the population size.

The *de jure* concept, was used as the basis for the 'modern' censuses. In this concept the population size is based on the number of people who are supposed to live in a certain municipality according to the municipal registrars (Methorst 1902). Regardless if the persons where actually present during the count or not (KB[9] 1947). This included people who were temporarily absent, such as students, sailors, military persons, tradesmen etc. But also, persons who were living temporarily abroad but still registered at a municipality in the Netherlands were counted as part of the population. From 1879 onwards all the censuses based the population size on the 'de jure' concept.

The second concept, the actual or *de facto* population was briefly used for the censuses of 1849 and 1859-1869. In this concept every person who was present during the enumeration was counted as a resident of the given municipality, even though the actual residence could be in another municipality. Table 2.2 gives an overview of the different ways the population is counted over the years.

[9] KB 1947. Koninklijk Besluit. In: Staatsblad van 5 februari 1947, betreffende de twaalfde algemene volkstelling (met daaraan verbonden woningtelling). In: Staatsblad 1947, nr. H44

Censuses	Concept of Population
1829, 1839	'de jure'
1849	'de facto'
1859, 1869	'de jure' and 'de facto'
1879-1971	'de jure'
1981-2001	'de jure'

Table 2.2 – Overview of the different ways of counting the Dutch population (van Maarseveen 2002).

Administrative: Verifying the population register

The second objective of the census focuses on *administrative* purposes. This entails the construction, checking and updating of the population registrars based on the censuses. One of the objectives of the 1849 census was to form the basis for the *construction* of population registers in every municipality (Methorst 1902). From 1850 onwards all municipalities were obliged to maintain and *update* these records with current data from the census (van Maarseveen 2002, Verhoef 1981).

The municipalities used a family registrar system for the construction of the population registrars. This system was quite cluttered and cumbersome to use. Since these systems were based on the location of living, a move to another address implied that a whole family was crossed out and rewritten at another place or another volume. In cases of changes in the family, the pages in these registers were updated and annotated by way of crossing out old text and adding the latest information to it. Around the year 1900, large cities introduced the family card which improved the system to the extent that a change of address did not imply anymore a complete removal of the household to another page in the register.

The CCS (Centrale Commissie voor de Statistiek), already in 1896 and again in 1909 and 1916 proposed to no avail to introduce a 'personal card' system, to improve and replace the family registers. This new system would allow to record all major changes during a person's life, such as change in marital status and place of residence. Moreover, using personal cards would enable a more efficient check during census (as the data were collected on separate sheets). Although the government was more favorable

than before toward this idea, due to the Great War it was not applied in practice.

In 1919 the government had to make major cuts and was looking for ways to save money. With this in mind, the Minister of Interior asked the CCS if it was possible to eliminate the census after the introduction of the personal cards. Methorst (director of the CCS) rejected this proposal fiercely, arguing that the census was not only a valuable tool for checking the population registers but provided valuable insights into the population, groups and contexts (such as mortality tables, occupational statistics, important resource for research on housing conditions etc.) which could not be derived from the population registers[10]. He admitted that although some information could be extracted from the population registers, it would still need extensive manual checking. He showed the Minister for example that in 1919 around 80,000 people would be over-registered. Methorst in the end showed the minister that next to these obvious benefits also monetary benefits applied, leading to a withdrawal of the proposal to cease the census (van Maarseveen 2002). The discussions around this topic were not forgotten. In the end the introduction of the personal cards system did take place in 1938. This however was not related with the end of the historical census.

2.2.5 CENSUS CARETAKERS

Before the CBS was involved with the censuses, it was the responsibility of the Ministry of Interior Affairs to organize the censuses. The collection and processing of the data took place in

[10] CCS archive, Letter of January 1919

close cooperation with the provinces and municipalities. During this period the 'Statistics Bureau' of the Ministry was responsible for the organization and implementation (van Maarseveen 2002). In 1892 the CCS was founded and from that moment onwards the Ministry relied on the CCS for advice with regards to the Census.
The wishes and information needs of the municipalities and provinces were of great importance. In order to meet these requirements, after each census a feedback round was held in which the Provincial Governors reported back to the minister on how the census went. The content, rules and regulations, instructions for enumerators and municipalities were based on advice from the CBS and CCS. Building on this advice the Minister would issue the census as part of the Order in Council (Maatregel van Bestuur) in a Royal Decree. These documents were distributed top down and forwarded from the minister to the Provincial Governors, which in turn distributed this to the municipalities. While the municipalities were responsible for the census taking, the CBS did the checking. Currently the CBS is still actively involved as one of the caretakers of the historical Dutch censuses.

Whereas the CBS and CSS where mostly involved in the past with the content and checking of the censuses, other organizations have put great efforts into *preserving* and providing *greater access* to the censuses. The *Nederlands Instituut voor Wetenschappelijke Informatiediensten* (NIWI), now the institute for *Data Archiving and Networking Services* (DANS) worked together with the CBS in digitizing the censuses from 1795-1971. DANS is still highly involved in various projects aiming to provide greater access to the Dutch historical census data. The CEDAR project (see section

1.5) is the latest venture in making this data comparable across the years and enriching it by using Linked Data principles. All of the data, tools and output produced by CEDAR are currently archived at DANS. The expertise and care of the census by DANS has exceeded two decades of knowledge, has included different generations of census researchers and crossed different scientific disciplines.

2.3 TRANSFORMATION OF THE DUTCH CENSUSES

2.3.1 BACKGROUND

Census data are the most important and reliable quantitative data sources on de Dutch population for the nineteen and twentieth century, with such a broad geographical coverage. Throughout its life span the Dutch historical censuses underwent several transformations in order to make them better accessible (Doorn, Jonker and Vreugdenhil 2001). From being aggregated and directly published into *books* to being digitized to *images* and later transcribed into *Excel tables*, each represents the census at a specific stage of its life span.

Over twenty years ago, the volumes with the published aggregated Dutch historical census data (1795-1971) were scattered around the country. These valuable resources could be found at different places for example the CBS, University Libraries such as Leiden, Utrecht and Groningen, the International Institute of Social History (IISH) or even found at flea markets. The use and accessibility of these books was obviously quite limited. Moreover,

frequent usage of the books led to the deterioration of the physical quality of the books. Physical presence and cumbersome manual efforts were required in order to extract data from the census. In order to provide better access to the censuses different organizations and projects worked together, starting in 1997. From this year onwards the *Dutch Statistic*s (CBS) and DANS, worked together in digitizing the books with the aggregative results of the censuses. This cooperation resulted in a digitized and transcribed version of the census of 1899, published together with the 100th anniversary of the CBS. The project *Life courses in Context*[11], a collaboration between DANS and the IISH, financed by the large investment fund of NWO was responsible for producing the tables with the transcribed digitized census tables for the other census years.

2.3.2 DIGITIZATION PROCESS

The digitization of the Dutch historical censuses has a long history and underwent different transformations throughout the years. These transformations are characterized by different conversion approaches which will be explained in this section, i.e. *medium* and *content* conversion.

The outcomes of the original census forms were aggregated and published in several books, having different volumes for different years. These books are the closest thing we have to the original data as the census forms were not preserved (from 1849 onwards till 1960). As a first effort in both preserving and providing better access to the original census books, different digitization projects

[11] http://www.lifecoursesincontext.nl

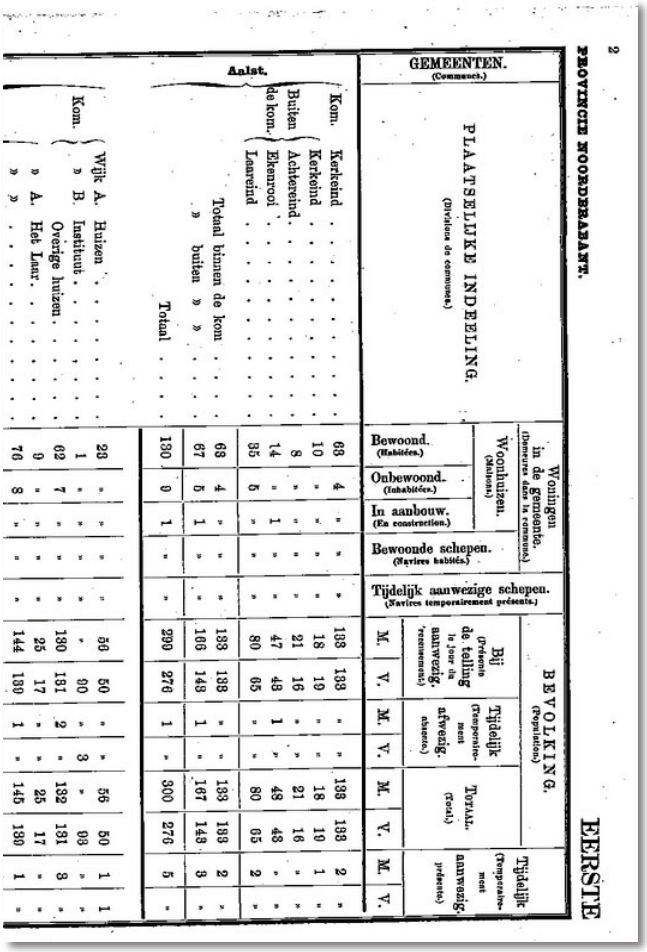
were undertaken by the CBS and DANS (Doorn, Jonker and Vreugdenhil 2001; Van Maarseveen 2008). These projects solely concentrated on the digitization of the census books and resulted in around 22,000 images. Although in principle no more than a *medium conversion,* this wealth of information was now made available in a much easier way through different platforms such as the internet and cd-roms. The *images* are the closest representation of the original books. However, although better accessible and preserved, the images as such are very difficult to handle. A single Table from the original census books can be represented by hundreds of images. This is especially true for census years such as 1899 where the tables are much larger. Moreover, the images are also quite unreadable on a normal screen without having them enlarged four or five times. Next to these issues the images are difficult to handle as the numbers in the images cannot be automatically processed and have to be manually transcribed. The digitized images have been systematically scanned and archived at the national digital archive (DANS) and available for all to consult.

The second step in the digitization of the historical censuses focused on *content conversion* to make the data more practical to use. As a result the images were *transcribed* into *Excel* workbooks. The conversion from images to Excel resulted in 2,249 Excel tables, representing the original structure and layout of the images. Experiments with Optical Character Recognition (OCR) did not lead to satisfying results; as a result the entire conversion was more or less done in a manual way. The main problem was that the automatic OCR conversion still needed extensive manual input such as checking and correcting *next to* the cost of digitization itself (Doorn, Jonker and Vreugdenhil 2001). Figures 2.1 and 2.2

show examples of respectively (a part of) an image of the census Table and the corresponding Excel Table after digitization. The images as well as the spreadsheets are downloadable and available online (www.volkstellingen.nl).

During the transcription of the Dutch census data 1795-1971 the choice was made to represent the data in a source-oriented way, meaning to resemble the content in the original sources as close as possible without making any interpretations. In the case of the Dutch census this went even further, by also trying to represent the *presentation* of the files in a source-oriented way. As we can see Figure 2.2 is an exact representation of the original Table as displayed in Figure 2.1. As a result researchers ended up with 2,249 heterogeneous Excel tables instead of one integrated dataset. Whilst this source-oriented digitization approach is typically a golden rule in constructing microdata, in the case of reproducing aggregate statistics, it can be a problem. This interpretation of the source-orientated method in the digitization process meant that no efforts were undertaken to harmonize the data and structure of the census tables. However, the source-oriented methodology as such does not require the visual presentation of data to be copied. This was a conscious choice made by the caretakers of the census (DANS) when transcribing the data.
As said, although easier to handle compared to the images, using the Excel representations of the aggregate historical Dutch censuses for research purposes is still a major challenge. Besides the data representation limitations of having over two thousand heterogeneous and unconnected Excel tables, another common problem relates to data harmonization. The Excel tables present many different variables and classification systems which need to be aligned in order to allow temporal comparisons (see next

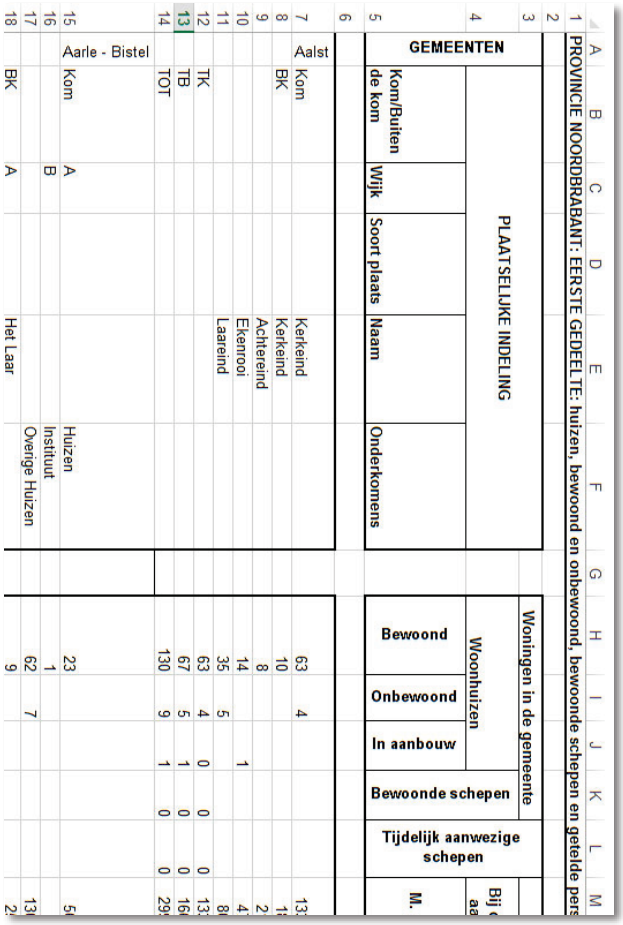
section). These Excel tables are the basis for the next steps in the digitization process of the Dutch historical censuses and form the starting point of our harmonization efforts.



2 PROVINCIE NOORDBRABANT. ERSTE

GEMEENTEN. (Communes.) PLAATSELIJKE INDEELING. (Division des communes.)		Woningen in de gemeente. (Demeures dans la commune.) Woonhuizen. (Maisons.) Bewoond. (Habitées.)	Onbewoond. (Inhabitées.)	In aanbouw. (En construction.)	Bewoonde schepen. (Navires habités.)	Tijdelijk aanwezige schepen. (Navires temporairement présents.)	BEVOLKING. (Population.) Bij de telling (Présents lors du recensement.) M.	V.	Tijdelijk afwezig. (Temporairement absents.) M.	V.	Totaal. (Total.) M.	V.	Tijdelijk aanwezig. (Temporairement présents.) M.	V.
Aalst. Kom.	Kerkeind	68	4	"	"	"	133	133	"	"	138	138	2	"
	Kerkeind	10	"	"	"	"	18	19	"	"	18	19	1	"
Buiten de kom.	Achtereind	8	"	"	"	"	21	16	"	"	21	16	"	"
	Ekenrooi	14	"	1	"	"	47	48	1	"	48	48	"	"
	Laareind	35	5	"	"	"	80	65	"	"	80	65	2	"
Totaal binnen de kom		68	4	"	"	"	133	133	"	"	133	133	2	"
" buiten " "		67	5	1	"	"	166	143	1	"	167	148	3	"
Totaal		130	9	1	"	"	299	276	1	"	300	276	5	"
Kom. Wijk A.	Huizen	28	"	"	"	"	50	50	"	"	56	50	1	1
" B.	Instituut	1	"	"	"	"	"	90	"	3	"	98	"	"
	Overige huizen	62	7	"	"	"	130	131	2	"	132	131	3	"
" A.	Het Laar	9	"	"	"	"	25	17	"	"	25	17	"	"
" "		70	8	"	"	"	144	139	1	"	145	130	1	"

Figure 2.1 - Example of a scanned image representing a Table from the original books



PROVINCIE NOORDBRABANT: EERSTE GEDEELTE: huizen, bewoond en onbewoond, bewoonde schepen en getelde pers

GEMEENTEN		PLAATSELIJKE INDELING				Woningen in de gemeente						
						Woonhuizen			Bewoonde schepen	Tijdelijk aanwezige schepen		
	Kom/Buiten de kom	Wijk	Soort plaats	Naam	Onderkomens	Bewoond	Onbewoond	In aanbouw		M.	Bij aa	
Aalst												
Aarle - Bistel												
Kom	A		Kerkeind	Het Laar	Huizen	23						5
	B		Kerkeind		Instituut	1						
			Achtereind		Overige Huizen	62	7					13
BK	A		Ekenrooi			9						2
			Laareind			14		1				4
TK						35	5		0	0		8
TB						63	4	0	0	0		13
TOT						67	5	1	0	0		16
						130	9	1	0	0		29
Kom			Kerkeind			63	4					13
BK			Kerkeind			10						1
			Achtereind			8						2
			Ekenrooi			14		1				4
			Laareind			35	5		0	0		8
TK						63	4	0	0	0		13
TB						67	5	1	0	0		16
TOT						130	9	1	0	0		29

Figure 2.2 - Example of a Table transcribed to Excel from images

Year	Tables	Annotations
1795	28	100
1830	17	71
1840	60	27
1849	94	75
1859	183	4896
1869	226	321
1879	985	516
1889	166	14349
1899	76	2594
1909	138	3381
1919	4	224
1920	48	5396
1930	32	1112
1947	133	83
1956	59	138
Totals	2249	33283

Table 2.3 - Distribution of the number of Tables and Annotations per census year

Table 2.3 shows the distribution of the in total 2,249 Excel *tables* and 33,283 *annotations* per census year (1795-1956). Each Excel Table applies to a certain year, specific region (municipal provincial and national level) and specific census type (population, occupation or housing census). Annotations may refer to annotations made in the census tabulations themselves or provide suggestions for corrections that were made during the conversion process into Excel. Given the source-oriented digitization approach the original figures in the tables were generally not

changed, even when the numbers were clearly wrong (although we found examples that indeed the source was corrected). We find annotations in different ways: as comments in a cell, in another sheet (i.e. Table) or even numbers which have been replaced without a description. Most of the annotations in the census are textual (whether a comment or interpretation) and only a small number are actual corrections of the data (numerical). All in all we deal with 33,283 annotations in the Excel files of which it is not possible to distinguish in a consistent way between changes from the source or 'new' annotations created during the conversion to Excel. About 40% of all the annotations are provided by the census of 1889 alone. As the process of annotating the census is still ongoing, and will continue in the future (the Excel files are still being checked manually for conversion errors), we need a flexible approach and generic solutions to allow future corrections on the data. In chapter 6 and 7 of this study we apply a specific model in RDF in order to organize the annotations, to deal with future changes in a consistent manner and to track back all corrections to the original sources.

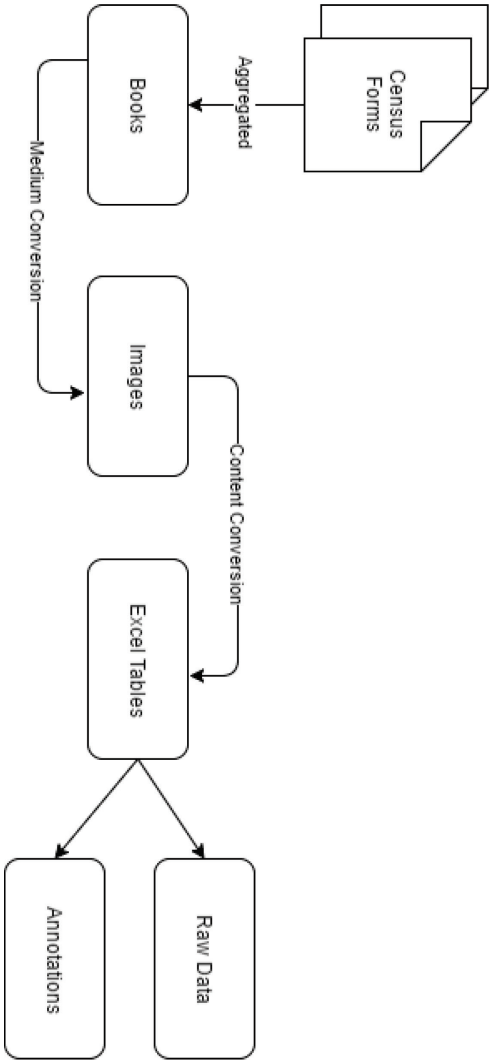


Figure 2.3 - Digitization Process of the Dutch Historical Censuses

Figure 2.3 shows the various stages of the digitization process from the census forms to the scanned images and Excel tables as we have described in this section. The final part is the integration of the Excel tables into an integrated database. This last step will be done by exploring the possibilities of Semantic Web technologies.

To conclude, during the conversion into different formats the source-oriented paradigm has had a reigning impact on the digitization process of the Dutch historical censuses. This conviction resulted in an uncompromising obedience in the digitization process, i.e. not only representing the data in a source-oriented manner but also the layout and structure of the data. We have shown that the Dutch historical censuses have underwent different transformations, each representing the census at a certain stage of its life span (see Figure 2.3). Providing this trail of transformations is crucial when harmonizing *historical* data as it not only adds more trust to our harmonization results but also provides a more comprehensive picture of the data in its entirety (i.e. where did the data come from, what are the original sources upon which the harmonization decisions are based etc.). When harmonizing aggregate historical data we need to able to *account* for the harmonized results, as this process often highly depends on expert decisions and interpretation. This means allowing the users to trace back our harmonization decisions and provide direct links to the original sources upon which they are based, both key aspects in historical research.

2.4 NEED FOR HARMONIZATION: PROBLEMS AND CHALLENGES

In general, the structure of a census is subject to change from year to year due to changing questions, variables, values, enumeration methods and classifications used. When dealing with historical statistical sources, especially census data which have been collected throughout different periods in history, researchers recognize the need for *harmonization* across the different sources as a fundamental requirement. The use of censuses which are collected over long periods of time has significant limitations because they are hampered with evolving variables, values, structures, observation methods, questions, processing methods and classification systems. This makes longitudinal studies difficult and prevents researchers to fully reap the potential of the census (Van Maarseveen 2008; Ruggles and Mennard 1995; Putte and Miles 2005; Ashkpour, Meroño-Peñuela and Mandemakers 2015). The Dutch historical censuses are not different and share many of these problems and even worse, providing only *aggregated data* in tabular form to work with.

In order to move towards a harmonized aggregate census database, we need to overcome the aforementioned challenges and enable the use of the census in a systematic and longitudinal way. In addition to the problems with the annotations discussed in the previous section (2.3.2), in the following sections we will describe the mentioned problems faced when harmonizing the Dutch censuses: working with aggregated data, changing variables and values, the need for creating variables and values, structural heterogeneity, inconsistencies and changing classifications.

2.4.1 AGGREGATE DATA

Statistical census data are typically presented on aggregated levels. This aggregation answers the information needs of the public, politicians, governments (as a decision making tool) etc. at given times. The specific harmonization challenge of the Dutch historical census relates to the fact that we only have aggregated data to work with. Although these type of data are not specific to the Dutch census (e.g. Belgium, United Kingdom, or the NHGIS project in the United States), in our particular case we aim to harmonize the aggregate data across all the census years, in comparison to current efforts which mainly focus on a per year harmonization of aggregated data.

Due to the lack of corresponding micro-data, harmonizing aggregated census data on a diachronic basis is hampered because it is not possible to simply build or rebuild a classification. Unlike many similar census harmonization efforts (see section 3.1) we cannot reconstruct the (classification) systems at a micro level to suit our needs. Our harmonization work therefore concentrates on two problems: First of all we need to harmonize the variables and values over time and secondly we have to harmonize the totals and subtotals from the several hierarchical layers in which the census results are published. The second problem arises e.g. when the national total of some specific variable is not the same as the sum of the provincial totals for that variable. The lack of micro data necessitates the use of a *combination* of statistical approaches with regards to harmonization of aggregated data. Considering this, we are constrained to provide higher level aggregations, create new variables and to use estimations, averages, ratios, interpolations, imputations and other methods necessary to

provide harmonized variables. This part of the harmonization process depends primarily on expert input and manual decisions, making accountability an important aspect of aggregate data harmonizations. In cases of estimations we would like to know who created it, which method was used, what kind of decisions and / or interpretations were made, upon which data their harmonizations are based on etc.

2.4.2 CHANGING VARIABLES, VALUES AND CLASSIFICATION SYSTEMS

Classifications systems are used in the census to categorize the various *variables* and *values* in order to put them into meaningful groups (Begthol 2010). Changes in the structure of the census and the evolution of the variables are also reflected in the different classification systems used in the Dutch historical census. Classification systems are not only meant to reduce the data to manageable proportions, they also provide researcher standards vocabularies to which they can refer. However, radical changes in the classification systems and how they are coded from one year to another make it difficult for researchers to utilize historical censuses for studies over time (Meyer and Osborne 2005; Pineo, Porter and McRoberts 1977; Ruggles and Menard 1995).

A general feature of the evolution of census variables over time concerns the l*evel of detail* provided. We find some variables which stay more or less the same over time, such as; gender, marital status, housing types etc., but with variations in labeling (including different spellings). In most cases the evolution of the census variables can be described through an *evolution tree* where we identify four different scenarios with regard to the changes the

variables undergo. A first scenario is the introduction of new variables (*creation*) reflecting the changing information needs at a given time. In other cases we find that certain variables were merely used for specific census years and removed from later censuses; we refer to this as *extinction*. Other common scenarios are the *merging* and *differentiation* (splitting) of variables throughout the census.

We encounter scenarios of creation, merging or differentiation often with geographical entities such as municipalities which have changed significantly over the course of time in the Netherlands. Censuses entail a rich geographical coverage of countries and their spatial division. Being able to connect data to a location enhances its meaning and makes allowances for their environment so as to enhance their comprehension (St-Hilaire et al. 2007). However, comparative studies over time and space are severely limited by the changing of boundaries. In the case of the Dutch census, the practical problem researchers deal with, is that the municipal boundaries in the history of the Netherlands experienced many changes (Van der Meer and Boonstra 2006). When municipalities merge, split, disappear or emerge we need uniform ways of accessing them both across time as well as space. In order to deal with this issue classification systems are created and used to provide standard codes for tracking the various municipalities over the years (Van der Meer and Boonstra 2006). To achieve a good coverage of all the geographical changes the classification systems have to be accompanied with a sophisticated system of shape files showing all municipal borders over time in order to do comparative studies (Manso and Wachowicz 2009; Owens et. al. 2009).

Examples of changing variables across time can also be found throughout the occupational census due to innovation (i.e. specialization, differentiation etc). One of the direct effects of industrialization was the enormous change that the labor market and, therefore, the occupational structure underwent (Baskerville 2010; Boonstra and Mandemakers 2004). Accordingly, throughout the years occupational categories of the census have evolved significantly. Categories emerged and disappeared over time (De Jonge 1966). This is a common characteristic throughout censuses worldwide. This process was partly due to the evolution of the occupation itself and technological reasons but also related with changing classification systems (Meyer and Osborne 2005). Accordingly, this evolution is reflected by the U.S Census of 1950 which defined 287 separate standardized occupations, compared to 441 in 1971. In the case of the Dutch census we see for example 327 occupational titles for 1849 divided by 36 classes. In 1899 (one of the most elaborate census years) we find 3900 occupational titles, 36 classes and four different social statuses given to each occupation. In 1947 (a downfall in the evolution of the Dutch census due to a very low budget), we have only 29 classes and no occupational titles.

During its life span the Dutch occupational census underwent several structural changes. Until the 1889 census a simple classification of occupations was used which counted all occupations into relatively broad categories without making any distinction in the kind of enterprise. After this period the occupational classification system changed significantly and recorded both the occupations as well as the kind of enterprise in which the individuals were working, providing a greater level of detail (van Maarseveen 2008). One of the features of this new

classification was that it also made a systematic difference between different types of hierarchical positions within an occupation/branch. The last three occupational censuses were less detailed and were combining an occupational census with a sector census, making separate categories for service employees within the industrial and agriculture sector. Accordingly, we can identify three different subsystems within the occupational classification system of the Dutch historical censuses: 1849-1859, 1889-1909 and 1920-1947.

Another example of values and classification systems which evolved significantly over time is religious denomination. The number of religious denominations has grown enormously over the years. The variable religion has already been introduced with the first census of 1830, but over the years there were many changes in the division of the denominations. While in some years there is a simple classification representing the most major religious denominations such as *Protestanten* in other years we have a very specific differentiation of religious types such as *Église National Suisse* or *Kwakers*. Extremely detailed examples are the fundamentalistic *Noorse kerk* in Almelo, *Zevende dag Adventisten* or the evangelic Moravische Broeders (Knippenberg, 1992). The classification of religious beliefs and phenomena has been a topic of study by many researchers since the 19th century (Müller 1873; Tiele 1897; Ward 1909; Jastrow 1914; Parrish 1941). Although various approaches and schemes were developed, they all have one thing in common, namely to bring order, structure and standards to such an unwieldy variable in research. What we need is a classification system allowing us to access the various religious denominations using standard values and variables. Examples of how these classifications are made are elaborated in chapter 7.

2.4.3 CREATING VARIABLES AND VALUES

The meanings of variables and concepts are subject to change from census to census. While in some cases it is simply a problem of different labels of a variable, we also find variations in variables which are much more difficult to harmonize. For example in the case of the 'housing type' classification of the census, we have very specific mentioning of how many people were counted in barracks, e.g. *Kazerne der Maréchaussée*, *Artilleriekazernes* or forts such as *Fort Isabelle* and *Fort Kijkduin*. As we don't have this detailed information for all years we combine these housing types according to the *function* they performed and create a new higher-level value called 'Military Buildings', belonging to the variable House Type. However, problems such as changing *age categories* require different statistical methods based on estimations. Variations in the classification of 'age ranges' in the aggregated results of the Dutch censuses make it difficult for researchers to use this kind of core data. As we deal with aggregated data rather than micro data, we cannot simply reconstruct new 'age ranges' to allow comparisons across time. New values need to be constructed to make age categories which cover all the census years. Because these new 'age ranges' are artificial (constructed by way of estimation, interpolation, imputation etc.) and made by domain experts with different restrictions and decisions in mind, we aim to provide *different* variables allowing researchers to choose a harmonization which fits their needs best.

The same flexible approach applies to the use of classification systems in harmonizing the census. As there is no one best solution, we need to provide the users different solutions for the same variable. In the case of the occupational census, DANS had

already connected the census to external classification systems such as the *Historical International Classification of Occupations* (HISCO) by Van Leeuwen, Maas and Miles (2002) in an early stage. However, the level of detail in the Dutch occupational census is much more fine-grained compared to the HISCO classification and using only the HISCO system would result in loss of detail. Next to using these types of external classification systems, it is also necessary to apply a bottom-up approach and use the classifications from the census itself to preserve the detail of the census. For instance, in the census we have occupations such as 'tile settlers and roof repairers' which were sometimes presented / grouped as one occupation and in other years separately. In HISCO they would both get the same code. Moreover, occupational titles as such may *not* change throughout the censuses but differences in content like the *changing social status* of an occupation makes it infeasible for historical comparisons when using HISCO. In the case of the Dutch census the occupation 'civil servant' is a good example of this. The status of this occupation has shifted a lot with regard to its original prestige and using a single HISCO code to refer to this function is not feasible. In this context we need to create standard classification systems for housing types, religious denominations, lower level classifications of occupations and of municipalities such as neighborhoods, area's etc. to preserve the detail of the census when harmonizing the data.

Although we are able to calculate various variables with a high level of accuracy, other variables are based on statistical computations. For example, in some years the population total is not given explicitly; however by adding the total males and females, we can reconstruct the 'population total' variable without

any doubt with regard to the validity of the harmonization. In other cases however, we need to perform more complicated calculations (estimates, interpolations, extrapolations, averages, imputations etc.) on the data in order to provide *at least one* harmonized version. This part of the harmonization process builds on manual input from domain specialists in which specific decisions and considerations are made. In some cases, simply adding up or dividing a category according to a certain ratio could suffice. This is, for example, the case of the *diamond workers* in the occupational census of 1899, in which they received their own category and were no longer grouped with *stonecutters* as in 1889 (see Figure 2.4).

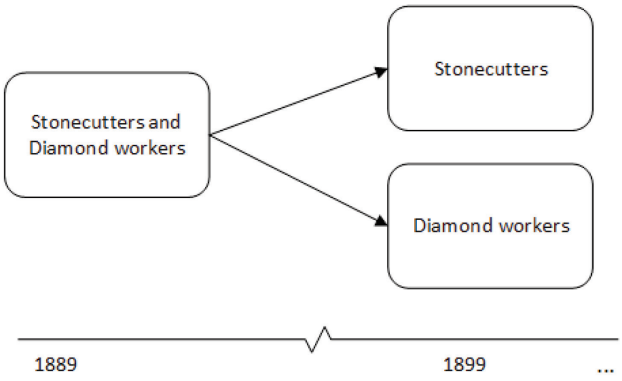


Figure 2.4 - Splitting of an occupational class

Accordingly, in this specific case we can provide two different harmonizations. On the one hand, we can combine the *stonecutters* and *diamond workers* of 1899 and create a higher-level variable for

comparison across the years, and on the other hand, we can split the occupational class of 1889 according to the ratio of *diamond workers* to *stonecutters* of 1899 and after.

We systematically keep track of all the changes and transformations made to the data so that users always know what has been corrected and where. By providing this provenance next to documentation on *variable* level users can judge among the differences and choose the most appropriate variables and harmonizations for their research.

2.4.4 STRUCTURAL HETEROGENEITY

During the digitization of the Dutch historical censuses the choice was made to apply a strict source-oriented approach and represent the images from the census books as closely as possible. Consequently, another harmonization problem of the Dutch census is related to the structural heterogeneity of the tables, even though the nature of the information in the tables is comparable. We therefore encounter not only changes in the naming and evolution of the variables, but also in the way they were presented, that is the structure (layout) of the tables. In order to move towards one system we have to determine how to model the different structures. While some tables have a basic structure of columns and rows with one or two levels of hierarchy, others introduce more complex structures. See Figure 2.5 for an example of two random tables from the census with distinct structures, presented as Excel tables.

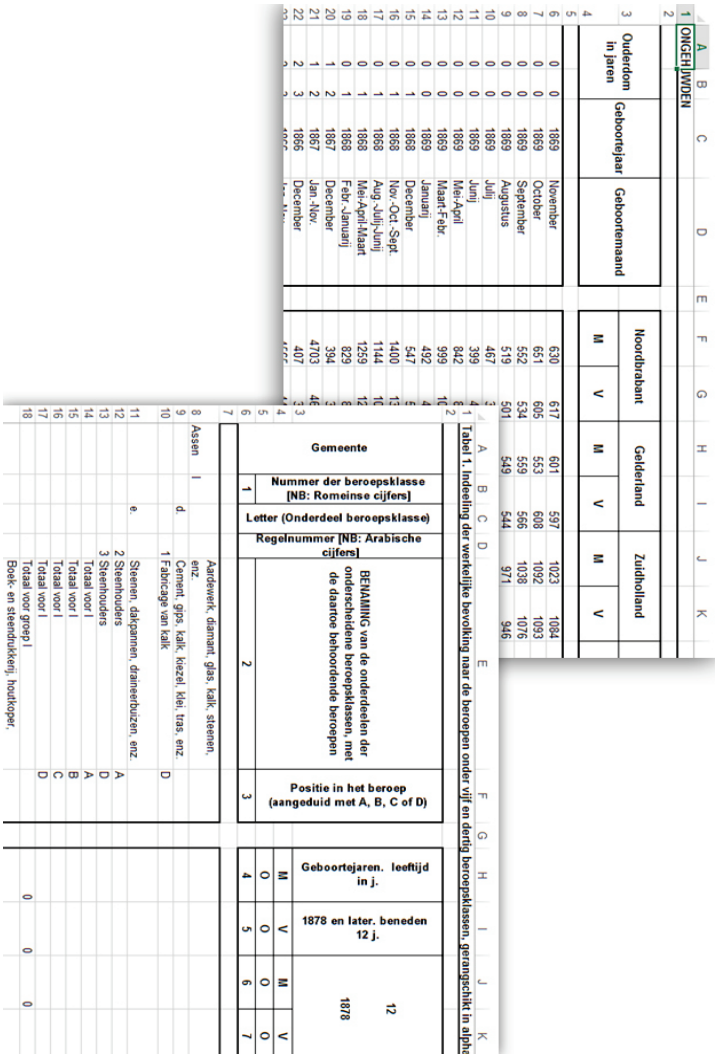


Figure 2.5 - Example of different Table structures

When building a database out of these different structures and hierarchies it becomes very difficult / impractical to find an overall model which would cover the entire dataset, without compromising valuable data. Trying to force an overall data model on the 2,249 Excel tables would practically mean that we have to

harmonize everything to the broadest category. Such a question (i.e. goal) oriented approach would result in the loss of valuable sub-categories or details which are only available for certain years. Preserving the heterogeneity of the tables is also an important research need from the perspective of historians and historical demographers which we aim to accommodate. Researchers interested in the original peculiarities of the tables must be able to retrieve any piece of data of interest from the original sources. Moreover, as we aim to provide several harmonizations for the same problem, we do not want to commit to a particular model when converting the Excel tables to RDF.

2.4.5 DEALING WITH INCONSISTENCIES

It will be clear that besides changing variables, classification systems and value meanings, the structural heterogeneity of the tables and the aggregated character of the data in itself may cause major inconsistencies when making one system out of the several censuses. However, inconsistencies are also present throughout the different censuses as they were published. The process of converting the data in the original census books to Excel files has not only introduced new transcription errors but also replicated source errors. In practice, it is impossible to distinguish between the two (unless one compares the Excel Table to the original census book, page by page, to see whether changes have been made). Even the original census books as kept in the libraries have handwritten changes to the data as numbers have been corrected. It seems that these corrections were digitized into the Excel files by way of annotations. The same happened with published corrections and with established mistakes during data entry.

Therefore, inconsistencies are not only present in the *structure* of the Excel files but also in the *numbers* transcribed as aggregate data.

In order to deal with the inconsistencies in the Excel tables we need solutions to clean, correct, enrich, standardize and even restructure the data before we can do research with it. All these 'improvements' to the data must be part of the harmonization package and are sometimes even necessary before being able to continue in the harmonization process itself. Data preparation is often underestimated and can take up to 60% of the total work (Garijo et al. 2014). In the case of the Dutch historical censuses we find for example spelling mistakes and variants, contents of columns which have shifted to another column or wrongly merged due to digitization errors, e.g. we find housing types under the municipality column, municipalities under the occupation columns, merged columns of the lower level municipal areas etc. As no consistent logic is applied it is very difficult to extract the correct data without extensive manual input from the unstructured and unconnected Excel tables.

2.5 CONCLUSION

Throughout history censuses have played an important, but often different role. Whereas ancient censuses had a more intrusive and negative connotation to it from which only bad things could come (i.e. taxes or war), the 'modern' historical censuses were designed to answer societal information needs. In western nations the censuses started to thrive around the mid of the 19th century. Not only were these data used by governments anymore, also researchers started to use statistics to answer societal questions. Quetelet pioneered in showing the importance of statistics to study social phenomena and greatly advanced the cooperation and comparability of statistical data among nations. In the Netherlands the census goes back almost two hundred years and captures a rich source of information through the population, housing and occupational census. Topics that can be addressed using these censuses are questions with regards to demographic characteristics, occupational developments and industrialization, housing needs and shortage and many others. Throughout its lifespan the Dutch census underwent different transformations and have mostly followed a source-oriented approach. The first source we identify are the original books in which the aggregated micro data are published. In order to provide greater access to the censuses various digitization efforts were undertaken, leading us to the Excel tables which we started this research with. In order to do comparative study over time and space with the census we acknowledged several key challenges such as dealing with aggregate data, changing variables, values and classifications, the need for variable creation and dealing with inconsistencies in the structure and numbers. We propose harmonization as the solution

they require. These challenges are coupled with the aggregate nature of our data and the changing nature of the historical censuses in general. In the remainder of this research we explore how Semantic Web technologies allow us to deal with the identified challenges and present generic and practical harmonization methods in chapter 6 and 7. However, before we go into the practical solutions we first explore the different approaches when moving towards a historical census database (chapter 3) and look at the possibilities of applying Semantic Web technologies on historical data (chapters 4 and 5).

3. THE THEORY OF CENSUS DATA HARMONIZATION

In order to comprehend the scope of current harmonization efforts and the different methods applied, we first give an overview of several key international census harmonization projects in section 3.1. We describe the different approaches taken in this field from three perspectives, i.e. whether they use micro or aggregate data to start with, whether they harmonize the data across years or across regions (e.g. harmonizing various geographical regions for a *single* year) and finally whether they use historical or contemporary censuses. In the next section (3.2) we go into the two main methods applied in the field of historical research when harmonizing data: the source and goal-oriented approach. Based on the used methods and goals of these approaches, we then take a closer look at census data harmonization and provide a specific definition in order to make the concept of 'harmonization' more tangible in section 3.3.

This chapter is based on parts of the following work. (1) Original work providing a view on the current landscape of census harmonization efforts (3.1). (2) Ashkpour, A., Mandemakers, K., Boonstra, O. Source Oriented Harmonization of Aggregate Historical Census Data: a flexible and accountable approach in RDF. Historical Social Research / Historische Sozialforschung, 41(4), pp. 291-321. (2016). In this work I was the main author and the developer of the source-oriented harmonization workflow we present in this chapter. This paper contributes to this chapter by describing the main two approaches when moving towards historical databases (3.2). Next it introduces our view and definition of historical census data harmonization (3.3). Next to this subsection extra background information was added to sections 3.2.2 and 3.2.3 as original work.

" I think there's been too much a tendency in the past to see the only important user [of statistics] as being the government. I think we have to get away from that"

Sir Michael Scholar [12]

[12] Chair of the UK Statistics Board 2008 - https://www.stat.wisc.edu/quotes

3.1 HARMONIZATION PROJECTS – CENSUS DATABASES

In this section we describe several harmonization projects that provide greater access to census data. In order to get a better understanding of the different approaches and their cohesion we have ordered them into four groups. The first group are the projects associated with and known as the Integrated Public Use Microdata Series (IPUMS). This 'IPUMS Family' of projects consist of the most advanced and comprehensive census harmonization databases we currently have. The second group we have identified are projects similar to that of the IPUMS Family in the sense that they are also dealing with micro data. Next we describe projects that are based on aggregate data and finally we close with RDF related census data projects. In each group we describe the goals of these projects and classify them according to their main differences i.e.; 'micro vs. aggregate' data, 'longitudinal vs. across region' harmonization and 'historical vs. contemporary' census data. To give an immediate overview of the key aspects of each project, we start with the main keywords at the beginning of each description.

3.1.1 THE 'IPUMS FAMILY'

Keywords: Microdata, Relational Databases, longitudinal harmonization

The Integrated Public Use Microdata Series (IPUMS) is one of the most well-known and comprehensive census harmonization projects to date, developed by the Minnesota Population Center (MPC). This center is one of the largest developer and contributor of demographical data resources, aiming to facilitate the use of census data as time series (Ruggles et. al 2003). The original IPUMS-USA project resulted in several census project spin-offs such as IPUMS-International, NAPP (North Atlantic Population Project) and IPUMS-NHGIS. Next to these projects, IPUMS also inspired others in harmonizing their data across time and space, using the coding and standards used by the 'IPUMS family'. The SweCens and Mosaic project build on the work of IPUMS, where SweCens provides harmonized data for Sweden and Mosaic for (Central) Europe and the Ancient world.

Although dealing with a long historical period and different geographical areas, common goals are defined which each different IPUMS sample must fulfil. The main goals as stated by IPUMS are[13]:

- Collection, preservation and documentation of the data
- Harmonization of the data
- Dissemination of the data

A common strength and major advantage of the IPUMS projects

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

is related to the availability of micro data. Instead of aggregated summary data (i.e. counts of persons with certain characteristics), the IPUMS data contains all individual responses transcribed from the manuscripts. By providing micro data many of the harmonization challenges described earlier can be dealt with. This type of flexibility of micro data allows researchers to construct their own variables, classifications, age groups etc. Micro data also provides insights into the relationships between certain characteristics at an individual level (enabling more sophisticated multivariate analyses), which is not possible using only summary data.

The IPUMS USA[14] project draws samples of the US population from thirteen census years, enumerating 55 million Americans and spanning from 1850 to 1990 (Sobek and Ruggles 1999). For most years before 1970, a one percent random sample of the population has been used. The project has made the characteristics of the US population available in the form of a database through a web-based system. The integrated US censuses provide insights for researchers into topics such as industrialization, urbanization, family structures, immigration etc. In order to allow comparability over time, the samples are put into the same format and a consistent coding scheme is applied to all variables and values (i.e. standardization) across the census years (i.e. longitudinal harmonization). Acknowledging the need for comprehensive documentation, the database provides next to the harmonized data, a complete documentation of all variables across each census year including their comparability.

[14] https://usa.ipums.org/usa/

IPUMS-international[15] is another major effort by the MPC, which aims to harmonize and encourage the use of multiple census years outside of the US from 1950 onwards. Collecting micro data from various countries in Africa, Asia, Europe and Latin America, the IPUMS-international database includes in total 159 samples from 55 countries worldwide (McCaa 2006). Following the same principle, micro data is used and converted to a consistent coding format to allow studies that compare census results over the years. Again, to accompany the correct use of the data, comprehensive documentation is provided. As an interesting side note, this project started of as a rescue operation by Robert McCaa for recovering original census tapes and got expanded over the years.

NAPP - North Atlantic Population Project[16] is a collaborative project carried out at the Minnesota Population Center (MPC), University of Minnesota, in cooperation with several international partners. The main goals of the NAPP project are similar to those of the IPUMS project. NAPP builds on micro data from seven countries for the period 1703-1911 and entails a machine-readable database of the complete censuses of Canada (1881), Denmark (1787, 1801), Great Britain (1881, 1911), Norway (1801, 1865, 1900, 1910), Sweden (1890, 1900), the United States (1880) and Iceland (1703, 1729, 1801, 1901). Together, these eleven censuses contain the richest source of information on the population of the North Atlantic world in the late nineteenth century. Next to the complete counts the NAPP has also made available samples for Canada (1852, 1871, 1891, 1901), Great

[15] https://international.ipums.org/international/
[16] https://www.nappdata.org/napp/

Britain (1851), the German state of Mecklenburg-Schwerin (1819), Norway (1875), and the United States (1850, 1860, 1870, 1900, 1910). A perquisite for temporal comparisons and analysis of human behaviour is to apply consistent coding, record layouts and documentation across all censuses (NAPP 2001b). Within NAPP uniform codes are assigned across all censuses and relevant documentation is presented into a coherent form. An advantage of the codes applied within the NAPP system is that these codes are based on and compatible with the existing IPUMS series of U.S. census samples.

The data currently in the NAPP database is the basis for a long-term cooperation to reconstruct the population of the North Atlantic world from the mid-nineteenth century to the present (NAPP 2001a). To use these data for longitudinal analysis several censuses are already linked on an individual basis by way of record linkage techniques. The NAPP database comprises of samples of the U.S. that link 1880 to seven other census years. Moreover it links males and couples between the 1865, 1875, and 1900 censuses for Norway. Similarly it also contains samples for Great Britain that *link* males, females and couples between 1851 and 1881 (Ruggles 2006).

SweCens [17]

Keywords: Microdata, Relational Databases, Record linkage, longitudinal harmonization

Sweden has two of the most comprehensive population databases, based on the countries' population registries, going back to the 18th century. Swedish demographic data are so valuable due to the time period they cover and the level of detail and quality of the enumeration process (SweCens 2011). These population registers contain very fine-grained *individual level* data. The SweCens project aims to harmonize the Swedish 1890 census and to produce an improved encoding of the 1900 census. The encoding is applied according to NAPP principles. By having the Swedish census data in the same format as other international censuses, the first goal of the project was realized, namely: making the data more valuable by allowing comparable research across different countries.

Next, the project aims to test and evaluate different record linkage methods on the 1890 and 1900 censuses and to link these census data to different demographic databases in Sweden. The SweCens infrastructure builds on previous results with regards to record linkage from projects by the aforementioned MPC (Wisselgren et. al 2014). The record linkage methods developed at the MPC are primarily based on data from the U.S censuses (i.e. micro data practices). In order to link the different historical censuses, NAPP contributors have harmonized the layout of the records, aligned

[17] http://www.humfak.umu.se/english/research/project/?code=660

various coding schemes and documentation for the different censuses and assigned standard codes across all the censuses.

The SweCens research infrastructure promises to improve the possibilities for national and international research using the historical censuses. By building an integrated infrastructure the project aims to expand their data to allow encoding, linkage and publishing of Swedish censuses from 1860 onwards (Wisselgren et. al 2014). Together with its predecessor 'TABVERK' (a DB containing aggregate information about the population in all Swedish parishes for the period 1749-1859) the SweCens project provides researchers with statistics on national and parish level for over 150 years.

Mosaic[18] Project

Keywords: Microdata, RDB, longitudinal harmonization, preserving, collecting and distributing

The Mosaic project is an international harmonization effort, coordinated at the Max Planck Institute for Demographic Research in Rostock, Germany, and was initiated in 2011 (Szołtysek and Gruber 2015). The data which Mosaic builds on is provided by its partners, an international group of institutions from around 29 different countries. The overall sample size of the Mosaic project is 904.500, spanning the period 1764-1918 (Szołtysek and Gruber 2015). See Figure 3.1 for an overview of the different partnerships in the Mosaic project.

[18] http://www.censusmosaic.org/

Figure 3.1 Current Mosaic partners. Source: (Szołtysek and Gruber 2015), design S. Gruber

The aims of the Mosaic project are to collect, harmonize and distribute historical census *micro* data to build a detailed resource for the study of historical censuses and al their aspects. The Mosaic project intends to extend the collection and distribution of census micro data for regions with surviving data from *continental Europe* and the *ancient world* (MOSAIC 2011). Contrary to the 'world' of NAPP, micro data from these areas have only partially survived and are mostly scattered. Working together with a group of international partners and institutions MOSIAC consolidates all the survived pieces of census data. Consequently, in order to allow comparisons across time and space, the records are distributed in the same standard format as the 'IPUMS family'. Next to harmonized data from various countries, MOSAIC also provides

the shapefiles and geo-referenced files of historical European administrative boundaries. The Mosaic project fills the gap for the missing ('census-like') data in Europe which are not provided by IPUMS or NAPP.

NHGIS[19] – The National Historical Geographic Information System

Keywords: Aggregated data, GIS tools, web-based access

The National Historical Geographic Information System (NHGIS) project was initiated by the MPC in order to make historical aggregated census data and compatible boundary files available through a Geographic Information Systems (GIS). The data is provided in standard formats such as spreadsheets and is intended to be incorporated into existing tools for data curation and analysis (e.g. Excel, R, SPSS etc.). Their custom online dissemination system allows users to create their own personalized datasets for research purposes without the need for any programming languages. The main goal of this system is to allow the users to *find* the needed data and *extract* these files, all in one environment.

This ambitious project builds a GIS which incorporates all available aggregate census information from the U.S. between 1790- 2015 with GIS boundary files (Fitch and Ruggles 2003). This includes surviving machine-readable aggregate data as well as 'new' data transcribed from printed and manuscript sources. The availability of detailed boundaries for key geographical areas

[19] https://www.nhgis.org/

allows for the reconciliation of changes in census throughout space and time.

In order to make this possible the data needs to be harmonized by formatting them consistently, developing full standardized machine readable documentation, creating highly accurate historical electronic boundary files and creating a web tool for the dissemination of the data (Noble and Fitch 2006). This project is somewhat similar to CEDAR in the sense that they build on aggregate historical data across time and space. However, the big difference between the two is that the aggregate data in NHGIS is built from *micro* data produced in other MPC-projects such as the IPUMS-USA.

3.1.2 U.K MICRO DATA PROJECTS

Similarly to the 'IPUMS Family' we identify another group of projects dealing with harmonization of micro data. In this section we present census micro data projects which are initiated in the United Kingdom by various institutions. These projects primarily build on their *own* infrastructure to provide greater access to the historical censuses of the U.K and are therefore separated from the 'IPUMS family'.

I-CeM[20] - The Integrated Census Microdata

Keywords: Microdata, RDB, longitudinal harmonization, Great Britain 1855-1910

[20] https://www.essex.ac.uk/history/research/icem/

The Integrated Census Microdata (I-CeM) project, was a collaboration between the UK Data Archive and the Department of History at the University of Essex, to harmonize the historical censuses of Great Britain (i.e. England, Wales and Schotland) between 1851 and 1911 (I-CeM 2014). The project ran for three years and created an integrated and standardized dataset, using information on individuals for almost all British censuses for the aforementioned period. The I-CeM project starts with already existing historical census datasets which were created by commercial companies for genealogical research, and enriches them with the rest of the available data in the censuses. Doing so, researchers gain detailed information (e.g. demographic, socio economic, geographical data) about every resident for the given census years. According to I-CeM the dataset is one of the most important historical datasets with regards to economic, social and demographic history.

The data in I-CeM is based on manual transcription of the original data. Although measures were taken to ensure great accuracy during the transcription process, corrections were still needed to improve the quality of the data. Derived from the census tables, I-CeM has created machine readable tables of population counts by year and developed a standard enumeration geography (i.e. the geographic areas of the census) for Great Britain over time (I-CeM 2014). The standardised administrative geography together with the digitised population counts have been used to integrate the data received from the commercial partners.

In the I-CeM project, next to reformatting the data and other perquisites such as checking and cleaning, the data is harmonized

by developing standard coding schemes for numerous variables. These schemes facilitate the coding of the data by providing standard dictionaries and thesauri and standardized geographical boundaries to allow longitudinal analysis. To ensure compatibility with other (historical) international census harmonization projects, relevant international classification standards have been added as well, i.e. the Historical International Classification of Occupations (HISCO).

For dissemination purposes, an online interface with supporting documentation has been created which allows registered users to extract the data from the dataset. Next to this, users are also able to create their own datasets which they can export for further use. To ensure provenance of the data after these harmonization processes (standardization, coding etc.), the I-CeM project provides the original text and numerical strings as separate variables, allowing researches to consult the original transcripts when needed.

EEHCM[21] - Enhancing and Enriching Historic Census Microdata

Keywords: Microdata, RDB, longitudinal harmonization

EEHCM stands for the Enhancing and Enriching Historic Census Microdata Project (Wolters and Woollard 2014). This project aims to create samples of anonymized records of the British census from 1961-1981. The EEHCM project is a cooperation between the UK Data Archive, the Office for National Statistics (ONS), National Records of Scotland (NRS)

[21] https://www.ukdataservice.ac.uk/about-us/our-rd/historic-census-microdata

and the Northern Ireland Statistics Research Agency (NISRA). The main goal of the project is to collect and recover British Census data of 1961-1981 and to harmonize these data according the standards of 1991 and 2001.

This process is structured into two phases (EEHCM 2012). The first phase focuses on the restoration and preservation of the 1961, 1966, 1971 and 1981 censuses of Great Britain. Data on record-level is recovered from archived data stored by the relevant organizations. This process involves extracting and transforming all existing data files and applying standardized variable and value labels to the raw data. The second phase of this project was the specification and creation of the Samples of Anonymized Records (SARs) from these complete datasets. This phase delivers datasets which are harmonized in the same way as the existing SARs for 1991 and 2001. The process is done by way of a workflow consisting of six stages: 1. Preliminary assessment, 2. Extraction and Quality, 3. Metadata and Preservation, 4. Cleaning, 5. Sampling and Deriving and closing with 6. User documentation. The output of these efforts, i.e. harmonized data, are available upon request via the UK data service. By using a standard workflow for the harmonization of such data, the project allows future data to be incorporated using the same standards.

Census Support[22] (UK Data Service)

Keywords: Microdata and Aggregate data, RDB, longitudinal harmonization

The UK Data Census Support is a service of the UK Data Service, to provide access to, and support for, users interested in the population censuses of 1971 until 2011 (UK Data Service 2012). The data is provided by three government census agencies for all nations in the UK. The results are initially only available to academics. However, the goal is to make these data available to a wider range of users as the project continues. This project provides various types of access for different types of users, including the necessary documentation and support. Its aim is to provide ready to use and access census data, in user friendly ways. The underlying data of this project builds on both micro and aggregate census data for the years 1971, 1981, 1991, 2001 and 2011. Boundary (GIS shapefiles) *and* flow data is also provided to allow comparison over space and time. The UK Data Service provides two different but complementary web interfaces to its data, named *InFuse* and *Casweb*. Using these interfaces users are able to access and extract relevant data from the (aggregate) census statistics and a variety of linked datasets and services.

Interfaces:
The UK Censuses have historically produced a set of *predefined* tables. These predefined tables are still hosted in their initial web interface, namely; *Casweb*. This older interface, contains 1971, 1981, 1991 and 2001 Census aggregate data as well as 1991 and

[22] https://census.ukdataservice.ac.uk/

2001 Census boundary data. These large and complex datasets comprise counts of persons and households for various geographical units. The format of these predefined tables in Casweb was not of much use for the users. Mainly because usres were forced to go through many tables just to find out if they contained the variables of their interest. In addition users had to go through different documentation to find the definitions, as the data and descriptions were separated. Because the formats of the census tables were hampering the creation of greater access to the censuses, it was decided to reformat the historical census tables into a structure based on standards.

InFuse is the newest web interface service of the UK Census Support, providing access to aggregate data from the UK 2001 and 2011 censuses. The InFuse system currently also has aggregated data for England and Wales 2001 (InFuse2012a). This interface allows users to select data by topics instead of tables. The InFuse dataset contains a selection of the UK 2011 Census data down to local authority level[23], as well as England and Wales's 2011 census down to output area level (InFuse2012b). The InFuse interface aims to guide users in selecting aggregate census data and presents characteristics and areas of interest to the users (i.e. a guided variable and value selection tool). Using this new format allowed this project to create a more flexible interface and access to the data.

A notable challenge which this project faced when dealing with aggregate data is related to the structure / heterogeneity of the datasets and the decisions to be made when moving towards one

[23] http://bit.ly/2uhLnz6

system. This involves the discussion of the source-oriented vs. goal-oriented approach in historical research which we will address in section 3.2. In their approach of creating comparable census tables, the data was restructured to conform to standards.

3.1.3 AGGREGATE CENSUSES

Although scarce compared to micro data harmonization projects, we introduce an example of aggregate census data harmonization by looking at the Belgian census data project in this section.

HISSTAT[24] – Historische Statistieken België

Keywords: Historical censuses, aggregate data, across region harmonization

The HISSTAT project started in 2009. HISSTAT is an inter-university network, aiming to generate and develop a research infrastructure of historical statistics. The main goal of this network is to develop a central database bringing together the historical statistics of all Belgian municipalities from 1800 onwards. The project is designed to protect, utilize and make accessible a multitude of historical census data for diverse applications and for the public at large. The research infrastructure has developed software that enables the processing of digital statistics and the creation of historical maps (HISSTAT 2009). The data is connected to maps via a Historical Geographic Information System, called HISGIS. The data made available by this project is harmonized for the given years but not longitudinal (i.e. across

[24] http://www.hisstat.be/

years). Therefore, the problem of *change* over time (2.4.2) and the difficulties it presents was not the main focus in this project.

LOKSTAT[25]

The Historische Databank Lokale Statistieken (LOKSTAT) project originally called Quantitative Database of Belgian Municipalities, was created with the intention to make census material available for research. LOKSTAT is one of the practical applications created by the HISSTAT project to further progress access to historical censuses.

The project aims to provide a centralized database for the use of these statistics, allowing greater use and access to the data for more meaningful and in-depth research. The digital series provided by LOKSTAT were largely achieved by transcribing the original figures. The LOKSTAT project is not only limited to the use of historical statistical data. LOKSTAT is also a center of expertise for the statistical analysis of data from *local* sources. It includes quantitative and quantifiable sets that cover the entire Belgian territory including lower geographical levels, i.e. the municipalities.

The main part of the data in LOKSTAT largely consists of data from censuses of the population, occupation and industry, agriculture and trade. Additionally, LOKSTAT until now contains information relating to other sectors of society. In their efforts LOKSTAT currently mainly provides aggregated data for the population census of 1900. This project is a great example of

[25] http://www.lokstat.ugent.be/

providing greater access to historical census data, however the data is merely harmonized for this given year.

3.1.4 RDF AND CENSUS DATA STUDIES

Since the introduction of RDF to domains outside of its own realm (Computer Science), various efforts have been undertaken in order to expose social historical datasets in the Semantic Web. The censuses were no exception and different international projects have converted various censuses to RDF.

The 2000 U.S. Census

Keywords: micro data, contemporary censuses, across region harmonization

The 2000 U.S. Census (Tauberer 2007) was converted to RDF providing population statistics on various geographic levels. In this exploration a *specific* model is proposed in exposing the 2000 U.S. Census into RDF. Although not historical and harmonized only across the year itself, it deals with the challenge of finding an appropriate data model to represent census data in RDF. This approach was one of the first explorations in using RDF and census data. This dataset contains 1 billion triples (data points in the RDF database) and has information on various geographical levels such as: 3200 U.S. counties, 36000 towns and 16000 villages. The data is made available via a SPARQL endpoint which is a web address where users can query the database. To extract detailed statistics from this RDF database, users can download the raw census files from the U.S. Census Bureau (1% or 5% samples), use a script and a patch file provided by Tauberer and

run these files locally on their own machine. The focus was obviously not on providing generic methods to harmonize and convert such data, but rather to show the possibilities of such an approach. In his exploration Tauberer also experiments on how different datasets and sources can be easily *linked* when converted to RDF. For example, the U.S census contains data about different states (with senators for each state). In a previous effort Tauberer already created a RDF file representing all member of Congress. As both systems use the same identifiers to denote states, the linkage between the census and members of congress of each state becomes a straightforward task.

The Canadian Health Census

Keywords: micro data, contemporary censuses, longitudinal harmonization

In Canada an experiment has been done with the Health Census. This census has been republished in RDF in order to provide greater access to and usage of the data (Bukhari and Baker 2013). The Canadian health census uses RDF in order to promote greater interoperability which accordingly cannot be achieved with conventional data formats. By using a scalable and interoperable format such as RDF this project aims to make the data reusable across different platforms and allow faster decision making.

Notably, in this project RDF is mainly used as a *final* step to expose and integrate the data as Linked Data. The actual data processing / manipulation (e.g. cleaning, preparing, standardizing etc.) was done using conventional methods and tools. By modeling the data beforehand they built on the same principles of

approaches which use relational databases. A reason for this could be that RDF is still in its early stages and compared to current technologies there are far less tools to work with. The data produced in this project is published according to the Linked Open Data (LOD) schema and incorporated with well-known Semantic Web vocabularies to enhance the interoperability of the data. The data is disseminated by providing a SPARQL endpoint and an (RDF) explorer interface.

The Greek 2011 Population Census

Keywords: micro / aggregate data, contemporary censuses, across region harmonization

In the context of a national large scale project regarding the management of socio-demographic data in Greece, Petrou et al. (2014) have explored and applied LOD technologies to the Greek population census of 2011. A similar goal compared to other census RDF projects, is to publish 'traditional' datasets such as structured Excel tables, into RDF and allow easier access and use of the census by third parties. Its aim is to develop a platform within which the Greek census is converted, interlinked and available in a LOD format. A five stage (sequential) method is proposed to make this possible, starting with *modeling* of the data as the first stage of the process. In the second step called 'Data RDF-ization', the data is cleaned, attached to unique identifiers (URIs), mapped and exported to RDF. 'Data Interlinking' is the third step of this workflow and involves interlinking the RDF data to external sources. In the fourth step called 'Data Storage' the data is saved in an RDF database called a "Triple Store". Finally, the data is open for all to use through a query endpoint and by

providing dumps of the RDF generated output in the 'Data Publication' stage (to be processed locally). In this approach reuse of existing standard vocabularies and RDF conversion tools is advocated, however they also recognize the needs for custom (domain specific) vocabularies. Although less focused on the harmonization challenges, it does deal with some aspects of harmonization as they advocate the use of standard vocabularies and propose a specific workflow to clean, convert and publish their data.

The 2001 Spanish Census as Linked Data

Keywords: micro data, contemporary censuses, no harmonization

The 2001 Spanish Census project is another advocate of applying LOD technologies such as RDF to the census. Accordingly such an approach encourages the development of the open government philosophy (Fernández, Prieto and Gutiérrez 2011). Using micro-data (a 5% sampling) from the 2001 population census, they propose a particular workflow for converting the data into open formats. Doing so allows for greater discoverability, accessibility and integration of these data by using the standards of Linked Data. Following such an approach would allow users to create their own tools and applications on top of the data, discover new relationships, integrate it with other datasets etc. The three main principles in this project are: 1. Using a standard format which is flexible enough to allow different uses 2. following the principles of open data and 3. to follow the principles of Linked Data to interconnect various data. It aims to stimulate greater reuse of these data by allowing third parties to build their own applications

on top of raw RDF data. Next to using standard vocabularies already available in RDF, they acknowledge the need to create their own vocabularies as many variables are still lacking as Linked Data. The data produced in this project is made available by providing the micro data files, population figures and a basic querying system. This querying system allows the users to construct predefined or new tables on a set of variables. More than harmonization, this project focuses on *publishing* open data and providing others the opportunity to interact with it and create their own tools and questions.

3.1.5 OVERVIEW OF THE CURRENT LANDSCAPE

Projects	Micro / Aggregate	Type of Census	Harmonization	Technology
IPUMS USA	Micro	Historical	Longitudinal	RDB
IPUMS International	Micro	Historical	Longitudinal	RDB
NAPP	Micro	Historical	Longitudinal	RDB
I-CeM	Micro	Historical	Longitudinal	RDB
SweCens	Micro	Historical	Longitudinal	RDB
EEHCM	Micro	Historical	Longitudinal	RDB
Census Support UK Data Service	Micro / Aggregate	Historical	Longitudinal	RDB
Mosaic Project	Micro	Historical	Longitudinal	RDB
NHGIS	Micro / Aggregate	Historical	None	RDB
HISSTAT	Aggregate	Historical	Across year	RDB
LOKSTAT	Aggregate	Historical	Across year	RDB
U.S. RDF Census	Micro	Contemporary	Across year	RDF
Canadian RDF Census	Micro	Contemporary	Longitudinal	RDF
Greece RDF Census	Aggregate	Contemporary	Across year	RDF
Spanish RDF Census	Micro	Contemporary	Across year	RDF

Table 3.1 - Overview of different Harmonization Projects and their characteristics

Table 3.1 shows an overview of the census harmonization efforts discussed in this section. To conclude we can state that there are many projects and studies focusing on providing greater access to census data. What this actually means in practice differs from

project to project. When looking at the landscape of historical census data projects we find very ambitious and fruitful efforts which go beyond accessibility issues and provide infrastructures to publish, harmonize and disseminate the data in various ways. Other smaller projects and studies are more in the beginning stages and have prioritized the conversion and availability of the data at this stage of their development. The most advanced efforts in this field are currently projects using historical relational databases such as described in 3.1.1. From this group we see that the 'IPUMS family' is highly represented (IPUMS USA, IPUMS International, NAPP, NHGIS) but also SweCens and MOSIAC strongly build on standards of the MPC and IPUMS. By using the same standardizations and codes, comparability among these censuses is ensured. In Europe the I-CeM, EEHCM and UK Census data projects are among the most advanced micro data efforts. They provide their own infrastructure and harmonize historical census across time and space.

The HISSTAT and LOKSTAT project have set ambitious goals but are (at the time of this study) still in an early stage. The next set of projects we presented all focused on exposing census data to RDF (3.1.4). By converting the data to RDF these projects aim to gain the benefits of Linked Data, e.g. connecting with other relevant systems, allowing others to tap into the exposed RDF data and build their own tables, applications, queries etc. A common characteristic of these effort is the use of standard vocabularies wherever possible. However not all the variables of the census are already available in Linked Data, as identified by the Spanish Census to RDF project. Unfortunately, the common caveat of census to RDF efforts is that they do not deal with one of the most

important aspects in historical research, namely changes over time.

OVERVIEW ACCORDING TO THE IDENTFIED CRITERIA

When comparing Table 3.1 against the three criteria we identified it clearly shows the lack of projects focussing on *longitudinal aggregate historical* census data harmonization. As we can see from this overview, projects using traditional database methods (i.e. relational databases) mainly use *micro* and *historical* data as a point of take-off. In the case of the UK Data Service or data used in the NHGIS project also aggregate data is harmonized, but they still have access to the micro data. In the case of Linked Data / RDF efforts we find mainly micro data and some semi aggregate data approaches, however these are all based on *contemporary* census data. Moreover, the RDF census harmonization projects do not harmonize across time (which is the actual challenge), but only for specific census years. It is clear that thus far no research has been done into the possibilities of harmonizing *aggregate historical* censuses data in a *longitudinal* way using RDF/ Semantic Web technologies. In this study we look exactly at this problem and explore the harmonization of aggregate historical data.

3.2 SOURCE-ORIENTED AND GOAL-ORIENTED APPROACHES

Introduction to the problem

Digitization of historical sources for research and analysis purposes has been a common practice ever since the advent and increased usability of computers. Different institutions, such as libraries, government agencies, universities, archives but also individual researchers have been involved in the digitization of historical sources. Historians account for a great part of this digitization process as in many cases they have to make considerable efforts *themselves* in order to provide a digital version of a given source for analysis. Historians are therefore often the actual creators of the data and not so much re-users or even buyers of data.

Digitizing historical sources for research purposes often starts with finding an appropriate data model to represent the data. The way the data is transformed and how this relates with the (underlying) original source data are a central point of discussion when dealing with digitized historical sources. A large part of these discussions concerns whether the design of the models (databases) should reflect a historical situation as envisaged by the researcher, or whether it should reflect the structure of the sources themselves. A key part of the discussions refer to what level of digital representation is adequate for historical and cultural sources. In the field of *history*, this discussion has centered around the dichotomy 'source-oriented versus goal-oriented' modeling (Boonstra, Breure and Doorn 2004).

An important decision when moving from original sources

towards historical databases is at which stage, data modeling and transformations processes such as data cleaning, standardization, classification, statistical computations etc. need to be carried out. Next to this, the discussions do not only center on these data transformation practices but also how the original source data is preserved in this process. The change in meaning, concepts, definitions, contexts and how different entities are referred to throughout history changes a lot over time. Dealing with historical sources raises the problem of ambiguity. For example, place names are known to undergo major changes over time, which raises the problem to which geographical location a given place name actually refers to. In others cases historians deal with names or variants of names which require some sort of standardization in order to make comparisons over time possible. These decisions often depend on expert knowledge and input from researchers. *How* this input is translated into different views on the data (a key characteristic in historical research) is an important aspect of data harmonization. In this section we present the two main principles when creating historical databases, i.e. source and goal-oriented approaches.

3.2.1 THE SOURCE-ORIENTED APPROACH

Source-oriented modeling of historical databases is the preferred method in historical research. Being able to refer to the original sources and allowing different interpretations on the same data (at all times) is an important requirement in this field of research. Source-oriented data processing methods do not force the historian to make a decision on which methods to be applied at

the time of the database creation (Boonstra, Breure and Doorn 2004; Mandemakers and Dillon 2004; Thaller 1993).

" The utmost confusion is caused when people argue on different statistical data "

Sir Winston Churchill

According to Merry (2016) the source-oriented database is a digital replication of the original source data, recording every last piece of information. The model of the database is consequently a strict representation of the presentation of the data itself. Such an approach has been applied in the digitization of the Dutch historical censuses (Doorn and Maarseveen 2007). However, the data model can also be source-oriented while not being a strict representation of the data. The main principle of the source-oriented approach is that the *content* of the source can always be reconstructed, and not that it needs to be a mirror of the source as such. Source-oriented historical databases are not built with predefined research questions and goals in mind and are therefore more flexible. Such an approach is especially helpful as they allow us to convert the original historical sources as one to one copies in a database and let us assign the meaning *later*.

Digitizing historical sources for research requires certain decisions to be made in the early stages of the process, with regards to how the data is modeled. Researchers advocating the source-oriented approach recognize the challenges of working with historical materials and suggest to avoid processes such as standardization

during data entry. For example is "Smith" or "Baker" an occupation or a name? Many of these answers need further research by experts and it should therefore be possible to repeat the harmonization of the data with different rules. According to Thaller (1993), a source-oriented approach can be defined as:

"… Attempts to model the complete amount of information contained in an historical source on a computer: it tries to administer such sources for the widest variety of purposes feasible. While providing tools for different types of analysis, it does not force the historian at the time he or she creates a database, to decide already which methods shall be applied later" (Thaller 1993, p. 39).

Historical data should be handled as pieces of raw data when building database systems without making any assumptions about its meaning as this depend on interpretation given by researchers (Thaller 1993). The only meaning given is that which is already in the source itself. Therefore, the source-oriented database is not built with predefined research question in mind but aims to administer a wide variety of uses. Not making any assumption with regards to the meaning, also necessitates a process where the historian is not forced to make any decisions with regard to the *structure* of the model when creating a database. The central theme in the source-oriented approach is the separation of harmonization and the original data and preservation of the source in a way that allows it to be used for a diversity of research questions (Ashkpour, Meroño-Peñuela and Mandemakers 2015). According to Greenstein's (1989) definition, the source-oriented approach should allow two main requirements. Namely that, the same source is handled differently in various stages of historical research and that the uses of sources vary over time. In other

words, we need flexible system allowing changing/different interpretations on the same data.

The source-oriented approach produces databases containing fine grained data with all its peculiarities, waiting to be explored and defined in an iterative way, allowing different interpretations. Using current methods and following a strict source-oriented approach can be very expensive (time, money and energy wise). In the last part of this study (chapters 6 and 7) we introduce the possibilities of a source-oriented harmonization approach using RDF in more efficient ways.

3.2.2 GOAL ORIENTED APPROACH

The source-oriented approach is sometimes criticized in its attempt to try and represent the original as close as possible. Some researchers worry that the use of this approach by 'purists' is for the wrong reasons, namely due to the lack of priority for analysis and/or could have a misplaced faith in the authority of the text (Boonstra et al. 2006). According to the goal-oriented approach the value of the source itself can be put into perspective, as a mediated representation of the historical past (Denley 1994).

Before building a goal-oriented database it is necessary to know exactly what you want to do with the database and which questions are going to be asked (Merry 2016). When researchers aim to analyze a dataset with a particular and often predefined hypothesis to be tested in mind, the goal-oriented approach is considered a practical solution. Working with specific datasets and research questions sometimes requires a more direct approach

towards modeling the data. Especially if, for example transformations are needed in order to make the data usable, or if the quality of the source is questionable. According to Denley (1994) the goal-oriented approach is often used for mainly quantitative analysis purposes and applied by researchers with specific questions in mind, using regular sources, accepting some arbitrary decisions about data. Goal-oriented databases are much quicker to build and work with, however they make it difficult to deviate from the original purpose of the database (Merry 2016). The availability of standard and mainstream tools, limited time and budget to build a historically accurate dataset are some of the main reasons for using the goal-oriented approach (Boonstra, Breure and Doorn 2004).

Current harmonization projects lean more towards 'goal-oriented' methods, where the users of the data depend on and are bounded to choices and interpretations which have been set before (Cameron and Richardson 2005; Thaller 1993). Such practices and models are not designed in such a way to easily allow different views on the data. These, mostly micro data practices, result in only one version of a newly categorized and classified dataset therefore limiting the variety of research questions that can be answered.

3.2.3 THE NEED FOR A FLEXIBLE SOURCE-ORIENTED HARMONIZATION APPROACH

Even when designing source-oriented databases, compromises have to be made (e.g. do we represent peculiar and rare instances in the final harmonized database?) as no 'perfect' representation of

the source is possible (Merry 2016). In this research we study the possibilities and opportunities provided by RDF and aim to bring current efforts further by creating a truly source-oriented database design method for historical censuses. Applying Semantic Web technologies such as RDF to harmonize historical censuses, following a flexible and source-oriented harmonization approach is currently an unexplored terrain.

" What is history but a fable agreed upon? "

Napoleon Bonaparte

Current harmonization approaches mostly try to build one solution to fit an overall goal, i.e. only provide one interpretation / standardization for the variables and values. In order to deal with the problems of aggregate census data harmonization we need a more flexible approach which allows different interpretations on the data. Different *choices* and *compromises* have to be made in order to find a model to make the data comparable across years. Notably, sometimes one is forced to make such decisions due to the lack of quality and structure in some data sources but also often due to time and budget restraints. This approach, while sometimes a necessity, goes inherently and unintentionally against the main principals of the source-oriented approach.

For example, as we have seen in the previous section, even projects harmonizing census data using Semantic Web technologies such as RDF, still build on existing tools and methods to model, clean and link the data (even using relational tables in the process) in advance. These efforts use RDF mainly as the final stage to disseminate the data in the Semantic Web. Our aim is different in the sense that we aim to provide a holistic source-oriented

harmonization approach in RDF. Meaning, that we cover the entire spectrum of the harmonization process in RDF and not only use it as a dissemination technology.

When dealing with aggregate data, we noticed early on in the CEDAR project that finding an overall data model is even a more challenging task, mainly based upon the *interpretations* of expert users. During the initial harmonization experiments we found that different expert users provided different views to interpret the *same* data. Altough the suggested models shared many similarities, they were clearly two different interpretations. Following goal-oriented approaches when moving towards a harmonized database we would need to choose one single model to represent the original data (i.e. a single schema to which all rows of the tables conform). The problem with this approach is the fact that we have to make certain choices (e.g. which research questions could be asked in later stages) early on in the harmonization process for which we simply lack the knowledge a priori. Transferring more, or preferably all information of the original sources into a database (i.e. the source-oriented approach) provides maximum flexibility later on (Merry 2016). When working with historical censuses one deals with vast amounts of data which contain many peculiarities. This is even true for similar census tables and years. This problem necessitates the need for making bottom up (i.e. data driven) changes to the harmonization along the way as these peculiarities are discovered. This process of exploring the original data in order to gradually harmonize it is only possible when using a source-oriented approach. Building on this knowledge we recognize the need for an iterative and flexible harmonization approach. Such an approach is necessary to accompany the 'learning process' which is of great importance during the harmonization itself.

Moreover, by following the lines of the source-oriented paradigm we are also able to preserve the underlying links to the source data (the trail of transformations of the census), which is often overlooked in many harmonization projects. Doing so the harmonized results we provide are accompanied with a trail of links to the sources upon which the results are based. The introduction, removal and change of questions in the censuses across time reflects societal needs and information processing approaches at given times in history. So although these changes may not always seem rational, and certainly do not make the harmonization process easier, they are still significant for researchers who are interested in details of a given time and place in history. A source-oriented approach should therefore preserve and even enrich the harmonized database with the underlying data. In the case of the Dutch census these are the Excel files, the scanned images and original census books. When harmonizing *aggregate* data, we are bound to make even higher aggregations and estimations compared to micro data. This means that the harmonized variables which will be generated obviously have less detail compared to the original data. However, being able to still point and query for these detailed instances and ask not so common questions is still important and part of the discourse of historical research. For example this allows a local historian (not interested in the harmonized data) to look for a certain occupation in a small town in the province of Friesland for the year of 1889. He or she may even decide to use a different standardization to make it comparable with his own research and data. We aim to enable researchers to do this by following a source-oriented approach.

3.3 HARMONIZATION

Historical data harmonization, such as the unification of formats, structures and content of historical data, is a knowledge intensive task which highly depends on expert decisions and choices. Knowledge about the source data is an essential aspect of historical data harmonization (Mandemakers and Dillon 2004). Formal descriptions of the data can only be provided by advanced users of the data or those involved in the data creation itself. Expert knowledge about the source data and its underlying model is therefore essential in understanding the problems to be addressed during the harmonization process.

This holds even more true when harmonizing *historical* censuses overtime (Esteve and Sobek 2003). Harmonization is not a one-try process; it is an iterative process of trying and learning how the classified and harmonized data interacts with the original data. Currently, there are no clear definitions or guidelines explaining which steps need to be taken in order to make the data comparable over time. Even when users are interested in the *same* data their *motivations* and goals may diverge, meaning that *different interpretations* on the data are an essential aspect of historical database design (Greenstein 1989; Thaller 1993). To allow these different interpretations, it is essential to follow a source-oriented harmonization approach and not commit to a predefined standardization (i.e. interpretation) on the data. Although there are no clear guidelines, we can identify a *set of practices* which are currently applied by researchers in order to make census data comparable across time. These different practices *together* are what constitutes census data harmonization in our view.

In this section we deal with the so far 'illusive thing' called harmonization. The term census data harmonization has been loosely applied in different projects and studies. Sometimes it is just used a synonym for standardization, and in other cases it is a combination of specific data integration methods. When trying to integrate historical censuses, questions which have been lingering in the minds of many are related to: what is the actual practice of census harmonization? How can we avoid using different structures, categories and semantics etc. in order to ask a simple question across the years? Before giving a definition we need to understand the challenges historical census data harmonization entails. However, while current approaches lack a defined harmonization workflow and definition, the notorious problems and challenges we face have been thoroughly described in extant literature.

Building on current practices and harmonization efforts we identify the following four topics as key terms averting the harmonization of historical censuses: (1) integrating dissimilar data sources and formats, (2) dealing with changing variables, values, structures and classifications, (3) constructing a database which can be queried across the years and last but not least, (4) the existence of a practical and generic harmonization workflow. Taking these ('needs') into consideration we define source-oriented historical (census) data harmonization as:

" An accountable process of creating an unified and unambiguous version of a dataset, which is flexible enough to deal with the changing characteristics of the data, whilst not committing to a predefined interpretation, by gradually applying a combination of known harmonization practices "

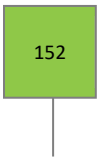
We adhere to the source-oriented approach of the digitization process of the Dutch historical censuses and extend this principle in our harmonization workflow. Our harmonization definition is accompanied by a practical workflow and technological backbone to support the preferred method of historical research (i.e. a source-oriented approach). Current efforts mostly provide, case specific harmonization and technical solutions to come to a final harmonized census database. In our efforts we focus on providing a *generic*, *structured* and *repeatable* workflow in order to make the harmonization process *itself* more explicit. We do this both in practical and technical terms, as we aim to be as transparent as possible and stimulate similar efforts on other datasets outside the realm of historical census data alone (Meroño-Peñuela et al. 2016).

3.4. CONCLUSION

In this chapter we have first presented some key historical census harmonization projects. When comparing these projects to the goals of this research, i.e. solving the problems of aggregate historical census data harmonization over the years, we find that none of the current efforts meet these goals. Projects working with traditional methods which are quite advanced but then again only deal with micro data. On the other hand, projects using RDF and census data are limited in the sense that they do not deal with a key concept in historical research, i.e. change over time *and* mostly converted *contemporary* censuses in RDF. What we learn from these efforts is that there is still a gap which needs to be filled (i.e. harmonization of aggregate census data across years). So, although the use of Linked Data technologies for census data publication and harmonization is not novel, our harmonization approach distinguishes itself in three different ways. First, we see *harmonization across time and space* as the most important step to make the data more usable, after publishing the data. Many current efforts merely aim to convert and publish historical datasets (such as census data) into RDF, with the anticipation of gaining Semantic Web benefits such as extending and enriching the data with other systems. However, mere conversion into RDF simply represents the data with all its faults and problems in another format. The harmonization part is usually absent in these practices, except in some cases where data were made comparable *for a given* census year by harmonizing over regions and levels of abstraction. Second, these efforts mostly use micro data as a point of take-off (whereas aggregate data provides different challenges). A third significant difference is that these projects harmonize

contemporary censuses and not *historical* ones, therefore dealing less with *changes* in variables, vocabularies, processing methods and classifications.

When looking at current harmonization efforts we basically find two different approaches to deal with these problems, i.e. source-oriented vs goal-oriented data harmonization. Current approaches lean mostly on goal-oriented harmonization methods, often due to budget and time constraints. However, as we explore in this study, the opportunities provided by RDF could eliminate many of these restrictions which hamper a strict source-oriented harmonization approach. This practically means that it is not needed to decide on how to model the data when moving towards a database system in advance, allowing different interpretations on the same data, preserving the richness of the original sources at all times and providing accountability for the harmonization actions taken during the data integration. As we have described (3.2.3) goal-oriented approaches are not suitable for solving the problems of the Dutch aggregate censuses. The challenges and problems our data presents, requires a flexible and data driven (bottom up) harmonization approach. Such a solution is best dealt with in the source-oriented approach. Interestingly, RDF and its characteristics present opportunities to allow a flexible and full-on source-oriented harmonization approach. In section 3.3 of this chapter we presented a clear definition of census data harmonization which we will use to build our workflow on (chapter 7). This definition is built on current efforts and presents key aspects which are related with solving the problem of aggregate census harmonization, i.e. flexibility, accountability and source-oriented principles.


152

Part II

HISTORICAL RESEARCH AND THE SEMANTIC WEB

4. SEMANTIC TECHNOLOGIES FOR HISTORICAL RESEARCH

In this study we look at how and if Semantic Web technologies such as RDF (i.e. Linked Data) are suitable for harmonizing aggregate historical data such as the census. The second part of this study on the other hand has a different perspective. In chapters 4 and 5 we look at key historical studies and how can benefit Semantic Web technologies and researchers. We survey and introduce various research efforts in the historical domain, to the realm of the Semantic Web and describe to what extent historical research can be addressed using Semantic technologies. The content of these chapter is based on a co-authored and cross-disciplinary work published in the Semantic Web Journal (SWJ) from the perspective of computer scientists. However, in order to tailor it more towards the audience of this book the article has been modified wherever needed. This practically means that certain text and technical terms are improved, edited or modified to make it better understandable for researchers which are not so familiar with the Semantic Web and its principals.

After a general introduction (4.1) of our survey on Semantic Web technologies for historical research we continue in 4.2 with introducing the Semantic Web and its principles. In section 4.3 we describe historical information science and how computer science has inspired historians from its early beginnings. In the following section (4.4) we describe the characteristics of historical data, its life cycle and the different ways in which such data usually are presented. We next describe the open challenges of historical data in section 4.5 and conclude with our findings in 4.6.

The second part, chapter 5, consists of a comprehensive overview of current work. This chapter revisits the open problems in historical data and historical research, and analyzes current contributions, namely papers, projects, online resources and tools, that apply semantic technologies to solve such problems. We study how successful these solutions have been and propose some challenges for the future.

Chapters 4 and 5 are a slightly adapted version of a jointly written article: Meroño-Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., van Harmelen, F. Semantic Technologies for Historical Research: A Survey. Semantic Web – Interoperability, Usability, Applicability, 6(6), pp. 539– 564. IOS Press. (2015). In this article I was responsible for describing the historical aspects of this study such as the work processes of historians, the information life cycle of historical data, the classifications of different type of data, description of relevant projects and contributed to the setup and collection of the survey itself.

4.1 INTRODUCTION

During the nineties, historians and computer scientists together created a research agenda around the life cycle of historical information (Boonstra, Breure and Doorn 2004). They also identified a number of problems and challenges in this field, some of them closely related to semantics and meaning. In this chapter we survey the joint work of historians and computer scientists in the use of Semantic (Web) methods and technologies in historical research.

Historians have a long tradition in using computers for their research (Boonstra, Breure and Doorn 2004, Thaller 1993). The field of historical research is currently undergoing major changes in its methodology, largely due to the advent and availability of high-quality digital data sources. More recently, the Web has introduced new research data publication methods, particularly since the inception of the Semantic Web (Berners-Lee, Hendler and Lassila 2001) and the Linked Data principles (Heath and Bizer 2011). This chapter looks into how Semantic Web technology could be applied to historical data, and whether and how these technologies can facilitate, boost and improve research in the historical field. Historical research is an interesting domain for the Semantic Web. Historical data are extremely context dependent, and always open to a variety of possible interpretations. Moreover, the availability of historical research data on the 'Web' is currently growing. The 'Semantic Web', is an evolution and extension of the existing 'Web'. The Semantic Web is based on the principles of *structured data* and *meaning*, in order to better enable computers and people to work in

cooperation (Berners-Lee, Hendler and Lassila 2001). In this chapter we study the crossroads of the Semantic Web and history as research domains.

We consider surveying the state of the art in Semantic Web and historical research applications a fundamental task for both fields. First, it is necessary as a knowledge organization task, in order to articulate research and determinte the contributions. Second, it fosters the development of semantic technology and history, both separately and as a unique field, and helps on building research agendas for both domains. Other attempts on gathering research efforts on Semantic Web and history exist, but most of them study specific history subfields (Pasin 2011; Segers et al. 2011; Kok and Wouters 2013) or analyze concrete task-oriented tools (Gangemi 2013; Robertson 2009a) and methodologies (Heath and Bizer 2011; Ide and Woolner 2004, 2007). Moreover, none of them consist of literature surveys or reviews. To the best of our knowledge, this is the first effort reviewing contributions on history and the Semantic Web as generic fields of research.

The elaboration of the study in this chapter is not free of obstacles. The first of them is the large amount of research contributions to collect and analyze, which had to be filtered to fit strictly the Semantic Web goals and the historical research goals of this survey. By historical research we mean strictly research performed by historians, and talk about history as a research domain. Thus, we exclude other fields of the humanities in which historical research is also performed, such as art history or history of literature. Nevertheless, in the end the number of contributions amounts to more than a hundred. Secondly, and even though the

corpus of available literature is large, we also encountered difficulties on accessing some of the sources (from articles to past research projects). To solve this, we conducted eight interviews with pioneers in this area and combined the contributions with the knowledge of domain experts. Thirdly, the clash of the vocabularies used by two different research communities, usually pointing at similar issues, is problematic. To bridge different jargon we devote some space to describe existing classifications of historical data, especially discussing terms like *structure*, and we map *historical data problems* in terms of *Semantic Web solutions*. While we concentrate on historical research, similar solutions emerge also in other humanities fields at the turn to e-humanities or Digital Humanities (Berry 2012, Schreibman, Siemens and Unsworth 2004).

4.2 THE SEMANTIC WEB

The advent of the Semantic Web poses new perspectives, challenges and research opportunities for historical research. Envisioned in 2001 by Berners-Lee, Hendler and Lassila (2001), the Semantic Web was conceived as an evolution of the existing Web (based on the model of the document) into a Semantic Web, based on the paradigm of structured data and meaning. By that time, most of the contents of the Web were designed for humans to read, but not for computer programs to process meaningfully. Although computer programs could parse the source code of Web pages to extract layout information and text, computers and software had no straightforward mechanism to process the semantics. The Semantic Web aims to provide information well-defined meaning, in order to enable people but also computers

to work in greater cooperation (Berners-Lee, Hendler and Lassila 2001). More practically, the Semantic Web can be defined as the collaborative movement and the *set of standards* that pursue the realization of this vision.

The World Wide Web Consortium (W3C) is the leading international standards body, and the Resource Description Framework (RDF) is the basic layer in which the Semantic Web is built on. RDF is a set of W3C specifications designed as a metadata model. It is used as a conceptual description method: entities of the world are represented as nodes (e.g. "Dante Alighieri" or "The Divine Comedy"), while the relationships between these nodes are represented through edges that connect them (e.g. Dante Alighieri *wrote* The Divine Comedy). These statements about nodes and edges are expressed as *triples*. See Figure 4.1 for a graphical representation of a triple.

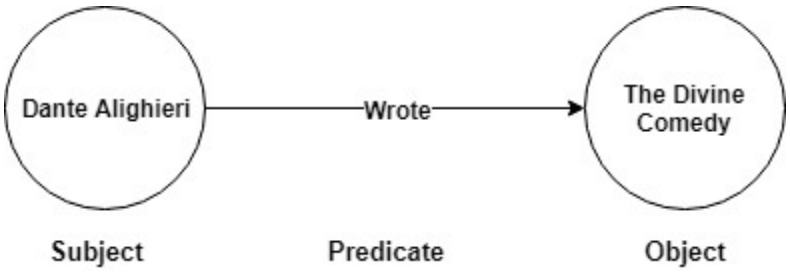


Figure 4.1 the 'triple' Dante Alighieri wrote The Divine Comedy presented as subject, predicate and object.

A triple consists of a *subject*, a *predicate*, and an *object*, and describes a fact in a very similar way as natural language sentences do (e.g. subject: *Dante Alighieri*; predicate: *wrote*; object: *The*

Divine Comedy). Subjects and predicates must be URIs (Uniform Resource Identifiers, the strings of characters used to identify and name a web resource like a web page does using URLs), while objects can be either URIs or literals, e.g. integer numbers or strings (Heath and Bizer 2011). RDF can be considered a knowledge representation paradigm where facts and the vocabularies used to describe them have the form of a graph. This setting makes RDF very suitable for data publishing and querying on the Web, especially when (a) the dataset does not follow a static schema (e.g. the changing nature of the census); and (b) there is an interest of linking the dataset to other datasets (e.g. by harmonizing 2,300 different tables across year).

Efforts on standardization have produced ontologies and vocabularies to describe multiple domains. An important note is that the term ontologies is sometimes coupled with ambiguity due to vast interdisciplinary collaborations within different fields of historical research and computer science. Whereas in socio historical research the term 'classification systems' and 'vocabularies' are the standard, in Computer Science 'concept schemes' and 'ontologies' are mostly used to refer to the same. An ontology is an "explicit specification of a conceptualization" (Gruber 1993, p. 1) and contains the classes, properties and individuals that characterize a given domain, such as history. In the Semantic Web, the design of ontologies is done using the Web Ontology Language called OWL[26]. OWL consists of several language variants built upon different modalities of Description Logics (Baader, Horrocks and Sattler 2005), *a family* of formal

[26] The World Wide Web Consortium (W3C). OWL Web Ontology Language Overview. http://www.w3.org/TR/ owl-features/.

knowledge representation languages. Such languages allow reasoning, that is, to extract or deduce consequences and new knowledge from the original.

A large number of RDF datasets have been published and interlinked on the Web, using these ontologies and vocabularies and following the Linked Data principles (Berners-Lee 2009). In the middle of the document-Web and the data-Web, formats and vocabularies for rich structured document markup (such as RDFa[27] or schema.org[28]) are enabling software agents to crawl semantics from web pages, bridging the gap between the Web for humans and the Web for machines. These efforts have evolved the Web into a global data space (Heath and Bizer 2011) where data can be queried e.g. using the SPARQL query language (SPARQL Protocol and RDF Query Language) called SPARQL[29]. Although the transition from the document-Web to the database-Web exists in the form of these standards and technologies, the simple idea of the Semantic Web remains largely unrealized (Shadbolt, Hall, Berners-Lee 2006).

[27] TheWorldWideWebConsortium(W3C). RDFa:RichStructured Data Markup for Web Documents. http://www. w3.org/TR/rdfa-primer/.
[28] Schema.org.
[29] The World Wide Web Consortium (W3C). SPARQL Query Language for RDF. http://www.w3.org/TR/ rdf-sparql-query/.

4.3 HISTORICAL INFORMATION SCIENCE AND RESEARCH

The field of historical research concerns the study and the understanding of our past. The field is currently undergoing major changes in its methodology, largely due to the advent of computers and the Web (Boonstra, Breure and Doorn 2004).

Computer science has inspired historians from the start. *History and computing or Humanities computing* were labels used before the inception of the Web (McCatry 2003). Many pioneers in computer aided historical analysis have a background both in history and in informatics, and reflected early on about the usefulness of computational and digital techniques for historical research (Boonstra, Breure and Doorn 2004). Ever since the advent of computing, historians have been using it in their research or teachings in one way or the other. The first revolution in the 1960s allowed researchers to harness the potential of computational techniques in order to analyze more data than had ever been possible before, enabling verification and comparisons of their research data but also giving more precision to their findings (Anderson 2008). However this was a marginal group within the historical research: in general, the usage of computers by humanists could be described as occasional (Feeney an Ross 1994). According to Boonstra, Breure and Doorn (2004) the emphasis was more on providing historians with the tools to do what they have always done, but now in a more effective and efficient way. Concretely:

- **databases and document management systems** facilitated the transition from historical documents to historical knowledge
- **statistical methods** were used predominantly for testing hypotheses, although with time these methods were more valued as a descriptive or exploratory tool than as an inductive method.
- **image management** aided historians to digitize, enrich, retrieve images and visualize data.

Although computing tools are currently embedded in the daily life of most researchers, the use of these tools has not revolutionized all sciences equally. Accordingly, history failed to acknowledge many of the tools computing had come up with (Boonstra, Breure and Doorn 2004). Instead of improving the quality of the work of historians and assisting them in their processes, software developed for historians often requires attending several summer schools (Bos 1995). Currently there are still many challenges and information problems in historical research. These difficulties mainly range from textual, linkage, structuring, interpretation, to visualization problems.

Despite these challenges, computing in history and in the broader sense the humanities, also brought some significant contributions in certain fields like linguistics (corpus annotations, text mining, historical thesauri etc..), archaeology (impossible without Geographic Information Systems (GIS) nowadays), and other fields using sources that have been digitized for historical (comparative) research and converted to databases (Boonstra, Breure and Doorn 2004). The use of electronic tools and media

is incredibly valuable and important for opening up various sources for research which would otherwise remain unused. These different sources contain rich information from various fields, which are often digital in nature in the form of databases, text corpora or images. These sources, in practice isolated databases, often contain a lot of semantics, but their data models were asynchronously designed, making them difficult to compare (i.e the problem of harmonization). So, while more and more sources are being digitized, more attention has to be given to the development of computational methods to process and analyze all these different types of information (Haslhofer et al. 2011).

A key issue for historians and other humanities researchers when dealing with historical data for comparative research concerns the lack of consistency and comparability across time and space, due to changing meanings, different interpretations of the same historical situations or processes, changing variables or changing classifications, etc. To deal with this issues harmonization is an essential aspect. In chapter 6 and 7 we go into the details of historical census data harmonization using RDF and present a source-oriented harmonization workflow.

Though not all research dreams materialized in the way initially envisioned (Kok and Wouters 2013), the inception of the Web allowed historians to aim for world-wide, large scale collaborations, especially in the area of economic and social history. This kind of web based cooperation allows to collect, distribute, annotate and analyze historical information all around the globe (Dormans and Kok 2010).

Changes in historical research are closely connected to the emergence of new scientific methods, and this co-evolution holds for decades and centuries. Statistics has influenced many fields including history, and paved the ground for quantitative studies (Kuczynski 1985). However, these kind of historical studies became more and more the domain of sociologists, economists and demographers than scientists educated as historians (Ruggles and Menard 1995). Late important changes such as Linked Data are consequences of recent technological trends connected to the emergence of the Web (Nentwich 2003) and the inception of Semantic Web technologies (Antoniou and van Harmelen 2004).

4.4 HISTORICAL DATA

Since the introduction of computers, historical research has produced high-quality digital resources (Boonstra, Breure and Doorn 2004). Historical datasets encompass texts, images, statistical tables and objects that contain information about events, people and processes throughout history. Converted, transcribed or born-digital, historical datasets are now analyzed at large scale and published on the Web. Their temporal perspective makes them valuable resources and interesting objects of study.

In this section we describe the ecosystem where historical information lives. First we introduce the life cycle of historical information, which is the framework we use to study the workflow of historical data. Next, we propose a classification of historical data depending on several factors. Finally, we revisit the traditional open problems of historical data. Some of these

problems have found solutions in current Semantic Web developments which we present in chapter 5.

4.4.1 THE LIFE CYCLE

The main object of study in historical research is historical information, and the various ways to create, design, enrich, edit, retrieve, analyze and present historical information with help of information technology. It is important to distinguish historical information from raw data in historical sources. These data are selected, edited, described, reorganized and published in some form, before they become part of the historian's body of scientific knowledge. We use the life cycle of historical information proposed by Boonstra, Breure and Doorn (2004) to study the workflow of historical information in historical research.

Historical objects go through distinct phases in historical research. In each phase, these objects are transformed in order to produce an outcome meeting specific historical requirements. The phases can be laid out as the workflow of a *historical information life cycle* (see Figure 4.2 on the next page). The phases, although sequentially presented, do not always have to be passed through in rigorous order; some can be skipped if necessary. The phases are also quite comparable with the practice in other fields of science. The life cycle of historical information consists of six phases:

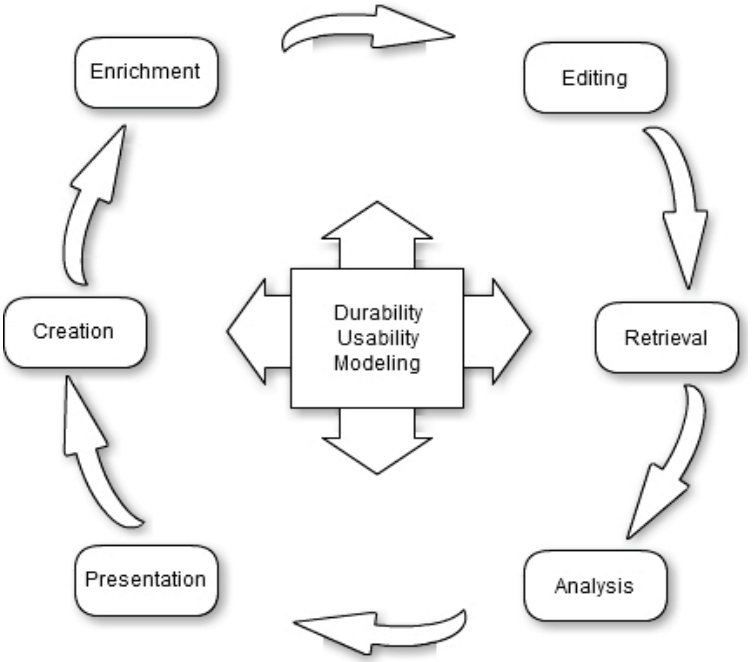


Figure 4.2 - Historical Information Life Cycle

1. Creation

The first stage of the life cycle is the creation stage. The main aspect of this stage consists of the physical creation of digital data, including the design of the information structure and the research project. Examples of activities in this phase would be the data entry plan, digitization of documents (through e.g. OCR), or considering the appropriate database software.

2. Enrichment

The main goal of this phase is to enrich the data created in the previous step with metadata, describing the historical information

in more detail, preferably using standards such as Dublin Core[30], and intelligible retrieval software. This phase also comprises the linkage of individual data that belongs together in the historical reality, because these data belong to the same person, place or event.

3. Editing

Editing includes the actual encoding of textual information, like inserting mark-up tags or entering data in the fields of database records, with the intention of changing or adding historical data of convenience. All data transformations through algorithmic processes prior to analysis also belong to this phase. Editing also extends to annotating original data with background information, bibliographical references and links to related passages.

4. Retrieval

In this phase information is retrieved, that is, selected, looked up, and used. The retrieval stage mainly involves selection mechanism look-ups such as SQL-queries for traditional databases or Xpath[31] and Xquery[32] for XML-encoded texts. The retrieval language for RDF is SPARQL. We use this to extract the data once converted to an RDF database.

5. Analysis

Analyzing information means quite different things in historical research. It varies from qualitative comparison and assessment of

[30] The Dublin Core Metadata Initiative (DCMI). http://www.dublincore.org/.
[31] The World Wide Web Consortium (W3C). XML Path Language (XPath). http://www.w3.org/TR/xpath/.
[32] The World Wide Web Consortium (W3C). XQuery 1.0: An XML Query Language (Second Edition). http://www. w3.org/TR/xquery/.

query results, to advanced statistical analysis of data sets.

6. Presentation

Historical information is to be communicated in different circumstances through multiple forms of presentation. It may take very different shapes, varying from electronic text editions, (online) databases, virtual exhibitions to (small-scale) visualizations. It can happen frequently in other phases as well.

In the middle of the historical information life cycle (Figure 4.2), three aspects are identified which are central to history and computing, but also in the humanities in general:

- *Durability* ensures the long term deployment of the produced historical information.
- *Usability* refers to the ease of efficiency, effectiveness and user satisfaction.
- *Modeling* denotes to more general modeling of research processes and historical information systems.

4.4.2 A CLASSIFICATION OF HISTORICAL DATA

The continuous usage of computing in different areas of historical research has produced digital historical data with different formats, perspectives and goals. To be used in the Semantic Web, these historical data have to be represented semantically, using the current standards of the Semantic Web (see section 5.2). In this section we propose a classification of historical data in order to bridge the gap between the data representation tradition in

historical research, and the standard modelling paradigms of the Semantic Web (Antoniou and van Harmelen 2004, Heath and Bizer 2011).

Primary and secondary sources

Historical sources can be characterized and divided in many ways. A basic distinction used by historians is between *primary* and *secondary* sources (Boonstra, Breure and Doorn 2004). Although we make this distinction we acknowledge that these terms can be used interchangeably and are not static notions, i.e. the secondary source of a researcher can be the primary source of another researchers.

Primary sources are original materials created at the time under study (Benjamins 2004). They present information in its original form, neither interpreted, condensed nor evaluated by other writers, and describe original thinking and data (Cook 2013). Examples of primary sources are scientific journal articles reporting experimental research results, persons with direct knowledge of a situation, government documents, legal documents (e.g. the Constitution of Canada), original manuscripts, diaries (e.g. the Diary of Anne Frank) and creative work. Primary sources can be distinguished into *administrative* sources and *narrative* sources, like biographies or chronicles. Administrative sources contain records of some administration (census data, birth, marriage and death rolls, administrative accounts of taxes and expenses, resolutions minutes of administrative bodies, deeds, contracts, etc.). Typically, historians want to extract these facts in order to gather statistical data. Narrative sources are full text documents containing a description of the past, made by an author being an eyewitness: think of

diaries, chronicles, newspaper articles, diplomatic reports, political pamphlets, etc. Historians may be interested in both, factual information (administrative sources) and the author's vision and the bias (narrative sources).

Secondary sources are materials that have been written by historians or their predecessors about the past (Tosh 2010). They describe, interpret, analyze and evaluate the primary sources. Usually, secondary sources gather modified, selected, or rearranged information of primary sources for a specific purpose or audience (Cook 2013). Examples of secondary sources are bibliographies, encyclopedias, books, review articles and literature reviews, or works of criticism and interpretation.

Since historical data have not been produced under the controlled conditions of an experiment, historical research always has something of the work of a detective, and certain details (read: annoying inconsistencies) cannot be destroyed or manipulated. These details may contain relevant information. On the other hand, to be able to extract statistical information and come up with more general statements, some formalization, relating information and harmonizing expressions of what is later used as variables is needed. Currently there is no common language to label these facts: the terminology is time and space bound, which makes formalization difficult and results dependent on different interpretations. Harmonization, the process of making data-sources uniformly accessible without altering its original form, is closely related to issues of standardization and formalization.

Intended further processing

Boonstra, Breure and Doorn (2004) propose to structure historical data depending on their type of required *further machine processing*. They distinguish between *textual data*, *quantitative data* and *visual data*. Textual data comprises the whole set of unstructured historical sources, such as letters, memoranda or biographies, all in a form of free text. Quantitative data can be seen as historical sources aiming at a quantitative analysis, like church registers, census tables and municipality register data. Finally, visual data gathers all kinds of historical evidence not encoded by text or numbers, such as photographs and audio visual sources.

Level of (data) structure

At the end of the creation phase (Figure 4.2) one may expect to have a historical dataset suitable for further processes. However, the nature of the steps to be taken thereafter may strongly depend on the way the resulting dataset is structured. Indeed, attaching Semantic Web technologies to these historical sources is strongly dependent on their level of structure. We propose the historical data classification shown in Figure 4.3 (see next page). Here we distinguish three levels of inner structure in historical datasets: *structured, semi-structured* and *unstructured*. Each level of structure can be divided into several *types of structure*.

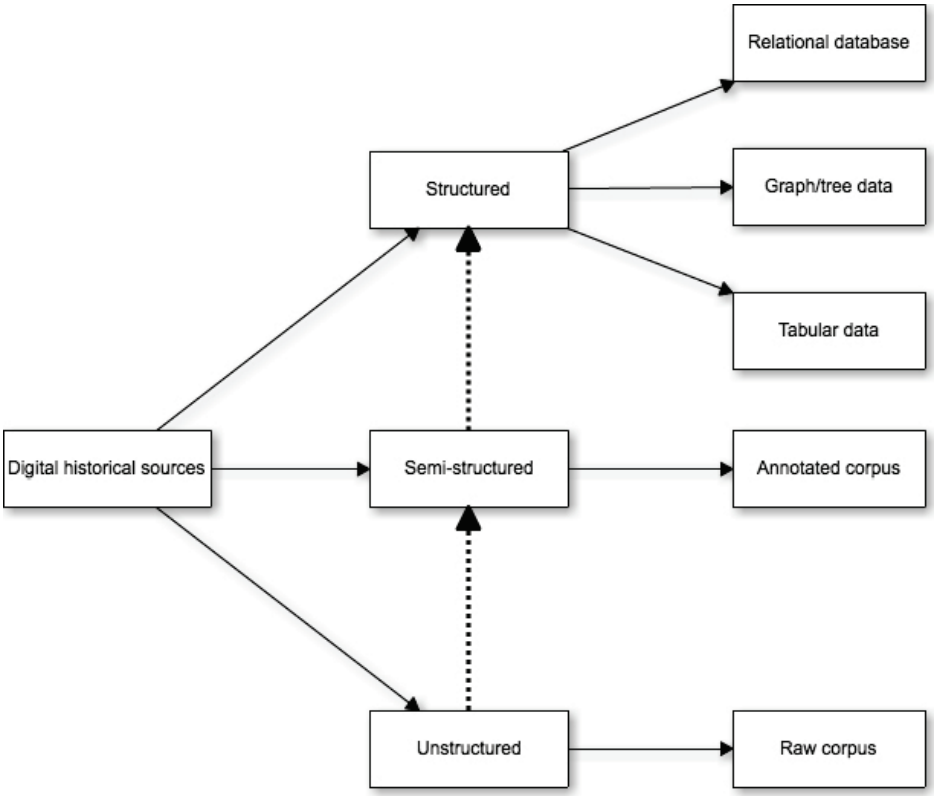


Figure 4.3 - Classification of historical data according to their level of structure

The dotted arrows in Figure 4.3 indicate the direction of usual transformations in workflows that identify historical entities (and their relations), from unstructured to structured representations. As this Figure illustrates, digital sources can be structured, semi-structured and unstructured. Moreover, the dotted lines show that one is able to go from unstructured data to semi-structured and finally to structured data.

Structured data

Structured data refers to sources that have a clearly defined data model. A data model is an abstract model that determines the structure of the data, organizes and standardizes data for communication. Data models are used as a 'plan' for developing databases and applications. An example of a structured dataset would be census material published in rows and columns, or a database of historical events. Well known generic examples of such a structure are sources encoded as relational databases, XML files, spreadsheet workbooks or lately even as RDF datasets. It is easy to see that all these examples meet a certain abstract model for the data they represent (relational schemas, DTD constraints, tabular formats and RDF triple statements). Structured historical data are usually managed with *relational databases*, *graph/tree representations* and *tabular representations*. Relational databases are the most well-known way of committing to some schema for representing historical objects and their relationships. Because their structure, relational databases are ideal for goal or model-oriented representation of historical data (Boonstra, Breure and Doorn 2004) with some concrete conception of reality in mind.

Relational databases

Relational databases have their own languages (SQL) and systems (MySQL, Microsoft SQL Server, PostgreSQL, Microsoft Access, Oracle, etc.) to represent and store (historical) data. They all follow the relational model (Codd 1969). Some issues, especially when trying to integrate data from different databases and modelled with different conceptual schemas, appear often in historical datasets encoded this way. These issues are for example

related to allowing different views on the data, providing a flexible data model, building source-oriented databases etc.

Graph/tree representations

Relying on graph theory, graph databases offer mechanisms for storage and retrieval of data with less constrained consistency models than (current) relational databases. They provide variable performance and scalability, high flexibility and complexity support. AllegroGraph, IBM DB2, OpenLink Virtuoso and OWLIM are typical examples. To exchange historical data in graph form, RDF (see Semantic Web section) is currently the mainstream model. Graph/tree data is found in historical samples that come in formats such as XML (trees), RDF (graphs) or JSON (JavaScript Object Notation). Although they are conceived for modeling data in very disparate models (a tree, a graph and nested dictionaries, respectively) and purposes (e.g. JSON is mainly used for data interchange between web applications and services), these formats also follow some assumptions to put structure on historical data.

Tabular representations

Some historical datasets are encoded in tabular form. Tables consist of an ordered set of rows and columns, the latter typically identified with a name. The intersection of a row and a column is a cell. Depending on the specific encoding (Comma-separated values (CSV), Microsoft Excel spreadsheets, etc.) tables can offer variable features. Tables are used to store all kinds of historical data, especially micro-, meso- and macrodata about individuals, registries, or historical population censuses.

Semi-structured data

Semi-structured data appear more often as an intermediate representation between unstructured and structured historical data than as raw historical data. Typical technologies applied here are markup languages, such as XML, to denote special characteristics of historical texts in specific regions of the corpus. *Annotated corpora* are the most important example of semi-structured data. They usually consist of raw historical texts with annotations on well-defined text regions, usually implemented with a markup language, like XML.

Unstructured data

In case a data model such as we described for structured data does not exist, we talk about unstructured data. In unstructured data there is scarce or no structure at all. The typical example is unconstrained, raw corpora encoded in plain text files. Unstructured sources are the most common representation of historical data, typically transcriptions of historical texts. Objects with a high variety of historical nature can be included in this category: letters, books, newspapers, memoranda, acts, etc.

Discussion

The use of the terms *structured* and *unstructured* in computer science to describe datasets is very different from the use of those notions in history, where administrative sources are often labeled as *structured* and the textual secondary sources as *unstructured*. Also narrative sources have internal structures, which can be made explicit. From the 19th century onwards historians have made scholarly source editions, which contain structured and annotated information. Nowadays the printed source editions are replaced

and supplemented by databases and XML-based digital editions. So, *structured* or *unstructured* are *relative* notions: administrative sources usually have an obvious structured layout, while narrative sources have a latent, at first sight *hidden* structure, which is made explicit as soon as they appear in a scholarly source edition. So, both administrative and narrative sources can appear in the form of *structured* or *unstructured* data in computer science jargon.

Although structure really matters for deciding what specific computing technique or semantic model has to be applied to the sources, being those sources administrative or narrative, deliberate or inadvertent, does not really matter if their inner structure is clearly identified. Their belonging to one type of another may have an influence at some point, but in general the procedure to extract RDF triples from the sources strongly relies on the type of source we have regarding their structure. The goal is a faithful representation of the source in Semantic Web formats: a source-close representation allowing to model data as a one on one copy, meeting the same requirements of faithfulness for critical source editions, which is the standard for historians. In alignment with this, in chapter 8 we present our source-oriented harmonization workflow. It is critical for semantic representations to consider *context* and *source structure* as critical editions do, because they may be relevant for interpretation of the data. A digitized, semantically-enabled historical source should ideally preserve context and structure and support source and goal-oriented extraction of data, in order to construct historical facts in the framework of a certain research. By means of dataset interlinking and appropriate design and usage of ontologies and vocabularies, *context* and *source structure* should be able to be preserved using semantic

technologies. To this end, ontologies / classification systems can be contextualized to conciliate a party's subjective view of a domain (Bouquet et al. 2004). By sharing knowledge intensive classifications systems such as HISCO for occupations or AMCO for Dutch municipalities in the Semantic Web, they can be disseminated and reused more easily (thus contributing to the acceleration of knowledge discovery in different fields of science).

Notably, the classification of historical data proposed in this section (according to their level of structure) is not strict and admits hybrid examples. For instance, annotated digital text sources can be provided both as XML files or stored in a relational database (e.g. for statistical analysis). Some authors classify sources that combine primary and secondary sources like these as tertiary[33] sources.

4.5 (OPEN) INFORMATION PROBLEMS AND CHALLENGES OF HISTORICAL DATA

Although many advances have been made in different fields of historical research and computers are seen as valuable assets, a high percentage of historians are unfamiliar with or remain unconvinced that semantic technologies may become a new methodological asset (Anderson 2008, Speck 1994). The reason is that the weapon of choice of historians was and remains mostly the database, particularly in relational form (Anderson 2008). This not only enabled historians to retain some of the integrity of

[33] University of Maryland. Primary, Secondary and Tertiary Sources. http://www.lib.umd.edu/ues/guides/

the original data sources, but also paved way for rapid advances on issues such as classifications and record linkage. The interplay between technological advances and historical research was not one which took place over night. Although computing and history was already introduced in the 1960's, it took quite some time before being accepted as a value adding new asset.

Historians typically do research using their *own* datasets, resulting in the creation of a vast amount of scattered data and specific technological challenges. In this section we revisit the traditional open problems of historical research derived from this tradition. We divide historical data problems into four main categories according to Boonstra, Breure and Doorn (2004): information problems of historical sources, information problems of relationship between sources, information problems in historical analysis, and information problems of the presentation of sources.

4.5.1 HISTORICAL SOURCES

The first set of open problems in historical research happens in phase 1 of the historical data life cycle (see section 4.4.1). This is when the historical data are created. Manually encoded or OCR-scanned, the creation of the dataset reveals the first barriers. Some characters, words or entire phrases in the original material may be lost or impossible to read or recognize by the human or the computer. Moreover, different techniques may extract historical entities differently. An example would be: what is the word that is written on this thirteenth-century manuscript? The next question usually is: what does it mean? *Background knowledge* is provided by libraries in the offline world. But the computer aiding tools also

need to have means to help the historian, using the Web as channel and semantics as meaning. Related to background knowledge is the provenance of the data. Even if the source is clearly identified and its meaning deciphered, the historian needs to know more. To which issue does it relate? What is the context? Why was it put there? Why was the text written? Who was the author? Who was supposed to read the manuscript? Why has it survived?

Another main issue relates to the structuring problem of historical data (Putte and Miles 2005). How can historical objects be encoded in a database? Researchers have to decide on what is an adequate *data model* for their datasets. As historians often have no clear research question when starting an investigation, it is neither possible nor desirable to model the data according to certain (i.e. predefined) requirements in advance. Moreover, different sources have been produced throughout different periods in history with different views and motives. Historical census data is a good example, having varying structures and changing levels of detail which hinders comparative social history research both in past and present efforts (Putte and Miles 2005). Especially when dealing with aggregate census data, the need for allowing different interpretations is needed. The main discussion regarding this involves whether to use a source or a goal-oriented data model for historical data (see section 3.2). Researchers in favor of the source-oriented approach claim that a commitment to a certain data model suitable for analysis should be postponed to the final stages of a project, in order to maintain flexibility and build on the data in a non-destructive manner. This is especially the case when the

database is supposed to be shared with other researchers or used in the future (Mandemakers and Dillon 2004).

4.5.2 RELATIONSHIPS BETWEEN SOURCES

As historical researchers deal with various isolated sources, they face the problem of how to *integrate* these dissimilar sources for their purposes. This typically happens in phase 2 (enrichment) of the life cycle of historical information (see section 4.4.1). An example would be: is this Lars Erikson, from this register, the same man as the Lars Eriksson from this other register? Does the label "Huizen" found in the Dutch census refer to Huizen (houses) or the municipality of Huizen, is "Mandemakers" an occupation or a name? etc. Thus, harmonization becomes a key aspect when historians try to ask questions across different sources.

Another example refers to micro data of the same person contained in different *censuses*, *parish registers*, *marriage* or *death certificates*. Obvious (record) linkage problems are how to disambiguate between persons with the same name, how to manage changing names (e.g. in case of marriage of a woman) and how to standardize spelling variations in the names. In databases, several issues affect data comparability. *Schema mismatch* occurs when two different databases cannot be compared because of semantic differences in the concepts of their defining schemas. For instance, two XML files conformant to different DTD schemas may define and structure differently the same historical entity. Additionally, *value mismatch* occurs when the allowed values for columns or variables in two databases are different. It may also happen across datasets despite being schema or vocabulary-

compatible. For instance, an attribute may encode the variable *social class* with categories A, B, C while other dataset may do so with categories *high, medium, low* or even worse, only *high and low*.

Other problems relate to how to link historical data with their spatial and temporal context. For example, some historical facts may need to be linked with occupational titles that evolve over time (HISCO[34]) or with countries with changing geographical boundaries. Compare for example the contemporary geographic position of countries in Europe with the situation in 1930 and in 1900; or the vastly changing boundaries of the Dutch municipalities (Van der Meer and Boonstra 2006). As historical research often deals with changes in time and space, historians require tools which enable them to deal with these aspects. Accordingly several techniques have been developed for historical research, but the applicability of these has yet to be determined (Boonstra, Breure and Doorn 2004).

4.5.3 HISTORICAL ANALYSIS

Historical analysis is a fundamental part of the life cycle (see phase 5 in section 4.4.1). It usually implies that data have been transformed and processed into datasets that are suitable for historical researchers. It also builds the bridge between their hypotheses and historical evidence.

The first issue in analysis is the massive treatment of historical data processed in previous stages to satisfy historical requirements, or

[34] HISCO Project. http://hisco.antenna.nl/.

to support a specific historical interpretation. An example would be: from this huge amount of digital records, is it possible to discern patterns that add to our knowledge of history? Various statistical techniques are borrowed from the social sciences to this end, like multilevel regression, and other techniques have been specifically developed for historical research, such as *event history analysis*. However, addressing historical data analysis in a broad sense remains essentially unsolved. Moreover, in historical research the meaning of data cannot exist without interpretations (Boonstra, Breure and Doorn 2004). Due to drifting concepts in history, different interpretations could exist with regards to certain data. However as interpretation of data is a subjective matter, this information should be added in a non-destructive way, preserving the original source data.

4.5.4 PRESENTATION

Presentation is the final phase of the historical information life cycle. Its goal is to use visualizations to aid the study and comprehension of historical data. An example problem of such phase would be: how do you put time-varying historical information on a historical map?

Presentation of historical data must be adequate. Different types of presentations are suitable at different stages of a research project. Presentation may take different shapes, varying from digitized documents, poorly and well modelled databases, or visualizations and representations on Geographic Information Systems (GIS). Currently there is a great need for tools and methods to present changes over time and space. In chapter 7 we

show how we used the outcome of the harmonization efforts and used e.g. NLGIS (an online resource containing the classification of municipalities and their corresponding shapefiles) to visualize these data across time and space using the historical boundaries.

4.6 CONCLUSION

In chapter 4 we presented the first section of Part 2 and looked at key historical research contributions which could benefit the development of Semantic Web technologies, in particular RDF (the basic layer upon which the Semantic Web is built). Chapters four makes two main contributions. First chapter 4, describes a classification of historical data depending on several factors, merging existing distinctions by historians with structural approaches from computer science. Second, it articulates the research conducted in the emerging field of historical Semantic Web and depicts the current landscape on advances in representing historical data with semantic technologies.

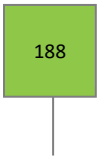
Throughout history different emerging methods and technologies were applied in the field of historical information science when working with (various types of) historical data. As our goal is to apply RDF technology on historical data such as the historical censuses we have first presented the Semantic Web and its main principles before we explore its possibilities. In this section we identified the Semantic Web as an effort aiming to provide well-defined meaning to various types of information, in order to enable people but also computers to work in greater cooperation. To realize this the Semantic Web pursues a collaborative

movement and uses a *set of standards* (vocabularies, classification systems etc.) which can be reused openly. RDF promises to be especially useful when the dataset suffers from structural heterogeneity and when one is interested with linking the dataset to other datasets.

We next looked at the interplay of computers and historical research and how historians have been inspired but also contributed to the development of new tools and methods from the early beginnings of historical information science. In our efforts to introduce research from the historical domain, to the realm of the Semantic Web, we want to explore to what extent this type of research can be conducted using such technologies. We therefore next looked at the characteristics of historical data and the challenges they present. To understand through which distinct phases historical data goes we have described the life cycle of historical information. This life cycle consists of six phases (creation, enrichment, editing, retrieval, analysis and presentation) which can be laid out as the workflow of a *historical information life cycle*.

The outcome of such a workflow produces historical data which can be characterized in many ways. The engagement of historians with computers in different areas of historical research has produced digital historical data with different formats, perspectives and goals throughout the years. In order to understand the diversity of historical data we have presented a *classification of historical data* in order to bridge the gap between the data representation tradition in historical research, and the standard modelling principles of the Semantic Web.

In the last section of this chapter we have revisited the traditional 'open problems' of historical research, specifically in combination with advances in the field of Computer Science. As we have presented, many technological advances have been made throughout history and historians were often at the forefront when exploring the applicability of such tools and methods. However, the adaptation of 'new' methods and technologies such as RDF in historical research is not self-evident and many researchers still have to be convinced by its use. Nonetheless, we believe that similar to the advent and use of relational databases which are currently deeply embedded in the workflows of historians, new methods such as RDF may become a methodological asset in historical research and beyond.


188

5. THE INTERPLAY OF HISTORICAL RESEARCH AND SEMANTIC WEB TECHNOLOGIES – FINDINGS: A COMPREHENSIVE OVERVIEW OF RELATED WORK

In this section we review the state of the art in the application of semantic technologies to historical research, describing relevant contributions towards a 'historical Semantic Web'. We look at contributions in the form of scientific papers, research projects, online resources (presentations, online articles), and tools, ontologies and lexical resources (demos, applications or programming libraries). Additionally, we map each contribution to one or more shared areas of concern. These tasks are shared areas of concern for both historical research and the Semantic Web. We start with historical knowledge modeling, presenting several contributions in historical ontologies and linkage with other data or systems. Next we look at the data integration issues and how Semantic Web technologies are used to deal with such issues. We close this chapter with open challenges and lessons learned.

5.1 HISTORICAL KNOWLEDGE MODELLING

In this section we study research that has been conducted to model historical knowledge using standard Semantic Web technologies. We group the contributions according to the emphasis of their research: *historical ontologies* and *linking historical data*. We also look at *text processing and mining* and *search and retrieval* to get a

comprehensive view of the different areas in which Semantic Web technologies could benefit historical knowledge modeling.

5.1.1 ONTOLOGIES

Data models are necessary for giving structure to any historical data, since they are the abstract models that document and organize data properly for communication. Ontologies encode such models in the Semantic Web (Berners-Lee et. al 2001), and attention has been given to the need of *historical* ontologies (Ide and Woolner 2007, Ashkpour, Meroño-Peñuela and Mandemakers 2015). In historical research, ontologies are the providers of metadata and background knowledge in phases 2 (enrichment) and 3 (editing) of the historical information life cycle.

We find a first category of such models in the form of (typically XML-encoded) taxonomies for historical research. A taxonomy (a specific type of ontology) is a collection of controlled vocabulary terms organized into a hierarchical structure. The first important example of such knowledge organization is the CLIO system, a databank oriented system for historians (Thaller 1980) which appeared already in 1980. CLIO included a tag/content representation for historical data that could be structured in complex hierarchies, supporting the recoding of material with uncertain semantics / meaning. CLIO remained as *the* system for organizing historical knowledge until the inception of the Web.

More recently, the Semantic Web for Family History[35] exposes a set of genealogy markup languages based on XML to semantically mark genealogical information on sources containing such historical data. In the context of the Text Encoding Initiative (TEI[36]) there is an important discussion about building the bridge between XML and ontologies in historical data. SIG: Ontologies[37] contains a full log on contributions on how to use ontologies with TEI formats. Namely, how TEI-XML encoded documents can refer to historical concepts and properties that have been previously formalized in an external ontology.

The Historical Event Markup and Linking Project (HEML[38], Robertson 2009a) was probably the first project with the goal of creating a Semantic Web of history. Started in 2001, it explored the use of W3C (World Wide Web consortium) markup technologies to encode and visualize historical events on the Web. Although in the beginning XML was the selected language to provide tagging and markup for describing historical events, the project later experimented with RDF to model and visualize them (Robertson 2009b). This transition was also happening in the whole historical ontologies community, as researchers better understood RDF and its differences with XML. One of the main difference was that XML is a syntax, whereas RDF is a data model with several syntaxes.

The modelling and representation of events, often defined as *persons* doing an *activity* in a certain *place* and *time*, has received a

[35] The Semantic Web for Family History. http://jay. askren.net/Projects/SemWeb/
[36] TEI (Text Encoding Initiative). http://www.tei-c. org/index.xm.
[37] SIG:Ontologies. http://wiki.tei-c.org/index.php/SIG:Ontologies
[38] HEML Project. http://heml.mta.ca/ heml-cocoon/description.

lot of attention in the development of historical ontologies, and most practical results show that the concept of the event is at the core of historical knowledge modelling. Van Hage et al. (2011) design the Simple Event Model (SEM[39]), intended to model events in the domains of history, cultural heritage, multimedia and geography. Similarly, the Event Ontology[40], inspired in the musical domain, models the representation of events as combinations of persons, places and moments in time. Finally, LODE[41]: An ontology for Linking Open Descriptions of Events is especially intended for the publication of historical events as Linked Data. Interestingly, these ontologies have a great overlap in their conceptual modelling of events even coming from different domains. On the other hand, some studies point out specific modelling needs for different historical domains. Stressing that, historical ontologies should reflect how a particular time frame influences the definitions of concepts (Ide and Woolner 2007).

Another big focus in historical ontologies is given to geographical modelling. Owens et al. (Owens et al. 2009) describe a geographically-integrated history, and stress the importance of dynamics and semantics in Geographic Information Systems (GIS). They set an agenda for historical GIS systems that includes important semantic modelling tasks involving ontologies and geography for historical analysis. Moot, Prévot and Retoré (2011) depict the interesting crossroad between text analysis, historical semantics and geography in a work that structures geographical

[39] SEM event model. http://www.cs.vu.nl/~guus/ papers/Hage11b.pdf
[40] The Event Ontology. http://motools.sourceforge.net/event
[41] LODE: An ontology for Linking Open Descriptions of Events. http://linkedevents.org/ontology/

knowledge from a historical corpus of itineraries. Vocabularies for historical place names are under discussion and development in the Semantic Web community[42]. Although not intended for historical research, the GeoNames[43] vocabulary is currently the reference for geographical modelling in the Semantic Web.

Since entities like places, persons or events change over history and time, there is work raising the importance of a change-aware modelling in ontologies (Flouris et. al 2008; Meroño-Peñuela et. al 2013). In historical research and the Semantic Web this is especially true for geographical names, places and regions (Hyvönen et. al 2011), but also for demographical, social and economic indicators such as occupations (HISCO). In the context of historical data such as the censuses (which change over time and space), domain specific ontologies / classification systems are needed for historians to fully engage with Semantic Web technologies. Such systems have been developed for many years by researchers in order organize and bring structure to the data, however the majority of them are not represented in the Semantic Web.

5.1.2 LINKING HISTORICAL DATA

By understanding the use and advantages of semantic technologies, practitioners and researchers of historical data can

[42] RDF vocabularies for historic place-names and relations between them. http://groups.google.com/group/ caa-semantic-sig/browse_thread/thread/ ae1db7fa31a1b5a0?pli=1

[43] The GeoNames Ontology. http://www.geonames.org/ontology/documentation.html

not only connect their own data sources internally (for example integrating the various historical censuses) but moreover, also disseminate their data into the Semantic Web. By doing so researchers can integrate their data with other data sources which were previously not possible or cumbersome. The approaches reviewed in this section match the historical data problem of the *relationships between sources*. As we will show, in most cases, the use of semantic technologies solves it.

If one side of knowledge modelling stresses the importance of ontologies and formalization of the semantics of historical domains, the other side pursues the usage of such ontologies to interlink related historical data on the Web. Some researchers in history have centered their interest in how semantics can help relating and linking historical sources and entities: "*historical, semantic networks are a computer-based method for working with historical data. Objects (e.g., people, places, events) can be entered into a database and connected to each other relationally. Both qualitative and quantitative research could profit from such an approach"* (Kalus 2007, p. 1). Linking historical datasets appropriately is an old and very well-known problem in historical research (Boonstra, Breure and Doorn 2004). The landscape on current projects linking historical data (typically extracted from unstructured sources) shows a tendency on publishing more and more historical Linked Data in RDF.

There is a wide variety of project types looking for semantic structure, though not doing so solely (or explicitly) in RDF. For instance, the Circulation of Knowledge and Learned Practices in

the 17th-century Dutch Republic (CKCC project[44]) studies the epistolary network for circulation of knowledge in Europe in the 17th century, extracting all entities and links from the correspondence of scientific scholars of that time. The LINKing System for historical family reconstruction (LINKS[45]) project reconstructs the links between individuals of historical families across several registries. The CCed[46] project follows a similar approach with clerical careers from the Church of England Database. While these projects mine the historical sources for important historical characters and their relationships, other approaches, such as the SAILS[47] project, dive into more concrete historical events and links various World War I naval registries together. The common goal in these initiatives is to produce a semantic network of historical data containing objects like people, places and events connected to each other, which clearly matches the intended purpose of historical ontologies (see ontologies section 5.1.1), but also the general mission of the Semantic Web (Berners-Lee et. al 2001) and Linked Data (Heath and Bizer 2011).

Many other projects expose their domain specific historical datasets using RDF. These datasets facilitate their linkage to others using existing ontologies, achieving shared goals with the old task of historical record linkage. For instance, the Agora project

[44] Circulation of Knowledge and Learned Practices in the 17th century Dutch Republic (CKCC Project). http://ckcc. huygens.knaw.nl/
[45] Links Project. http://www.iisg.nl/hsn/news/ links-project.php
[46] CCEd Project. http://www.theclergydatabase.
org.uk/publications/jeh_article.html
[47] SAILS Project. http://sailsproject.cerch.kcl. ac.uk/2010/07/about-the-sails-project/

(Agora[48]) aims at formally describing museum collections and linking their objects with historical context using the SEM (Simple Event Model). Historical events are found elsewhere in historical data. The FDR Pearl Harbor project[49] links events, persons, dates, and correspondence found on government letters and memoranda related with the Pearl Harbor attack on 1941 between the US and Japanese governments. All these entities are represented in RDF to model a graph of historical knowledge about that particular event. From a more socio-historical point of view, the 'Verrijkt Koninkrijk' project links RDF concepts found on a structured version of de Jong's studies on *pillarization* of the Dutch society after the World War II. More focused on media, the Poli Media project (Polimedia[50]) mines the minutes of the parliamentary debates in the Netherlands to link historical entities to the archives of historical newspapers, radio bulletins and television programs. The goal is to create a unified historical search environment, facilitating a cross-media analysis (Kemman and Kleppe 2013).

Some general purpose tools facilitate the creation of historical Linked Data. The Fawcett[51] toolkit and the Armadillo project (Armadillo[52]) are good examples. The latter exports RDF from any unstructured historical source, producing an RDF graph of historical knowledge that encodes the historical entities and their

[48] Agora Project. http://agora.cs.vu.nl.
[49] FDR Pearl Harbor Project. http://www.fdrlibrary.marist.edu/
[50] Polimedia Project. http://www.polimedia.nl/
[51] Fawcett: A Toolkit to Begin an Historical Semantic Web http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/175/217
[52] Armadillo: Historical Data Mining Project. http://www.hrionline.ac.uk/armadillo/armadillo.html

relationships expressed in that source. Other tools like Open Refine (OpenRefine53) or TabLinker (TabLinker54) are tailored to produce such Linked Data from structured sources such as tables and census data (see 5.4.2 a classification of historical data).

5.1.3 TEXT PROCESSING AND MINING

Although outside of the scope of our study, text processing and mining is another key area where historical research and computing meet in the realm of e-humanities. In this section we review work that deals with processing unstructured text. Textual resources play an important role in history research. We especially survey work on automatically extracting historical entities (such events or persons) via Natural Language Processing (NLP) techniques. The purpose of NLP is to enable computers to derive meaning from human, natural, or unstructured language input (see 5.4.2).

Structuring historical information from textual resources for further analysis is the bottom line of many research projects. The interesting differences come usually from the diverse source materials these projects mine. The general public-aimed Agora project enriches museum collections with historical knowledge in order to help users place museum objects in their historical contexts. To this end, Agora employs information extraction techniques from statistical natural language processing to extract named entities (actors, locations, times, event names) from textual

[53] OpenRefine. https://github.com/OpenRefine/ OpenRefine
[54] TabLinker. https://github.com/ Data2Semantics/TabLinker/

resources such as Wikipedia and collection catalogues which are used to populate SEM. Also, from the object descriptions, relevant historical entities are extracted which can be linked to the events. To formalize this workflow, Segers et al. (2011) present a prototype extraction pipeline for extracting events and their properties from text. They use off-the-shelf natural language processing tools such as named entity recognition and pattern-based approaches. The main problem they encounter is that the notion of events is still ill-defined in NLP research, and as such these tools are not yet readily available.

Textual encoding of the media have also been *the source* to extract historical knowledge in several projects. The Bridge[55] project aims at bringing more cohesion into Dutch television archives by finding relevant links between the official archives maintained at the Netherlands Institute for Sound and Vision and other information sources such as program guides and websites of broadcasting organizations. It is thus focused on improving access to television archives for media professionals. In order to do so, relevant entities are extracted from archives by using statistical NLP techniques. Furthermore, they identify interesting events in television archives by detecting redundant stories, utilizing the structure of the archive to identify links between different entities (Bron, Huurnink and de Rijke 2011). The Poli Media project mines the text of minutes of the general state debates to extract and link historical entities from the archives of historical newspapers, radio bulletins and television programs.

[55] BRIDGE Project. http://ilps.science.uva.nl/ node/735

The Historical Timeline Mining and Extraction (HiTiME[56]) project is aimed at detecting and structuring biographical events. To this end they analyze biographies of persons from the Dutch union history to create timelines that tell the life story of these persons, and social networks of the persons they interacted with. Van de Camp and Van den Bosch (2011) describe an approach to build networks of historical persons by mining biographies for person names and relationships between persons. They use standard named entity recognition tools and utilize the inherent structure of biographies (the topic of the biography is a particular person, and any persons mentioned in this biography should have something to do with this person) to detect interpersonal relations.

Many e-humanities and e-history projects are exploring document summary techniques or document enrichment techniques from NLP to aid search in their archives. One of these techniques is topic modelling[57], which can be used to add topic indicators to a document, which may help cluster search results or create more fine grained indexes of archive records. Wittek and Ravenek (2011) explore the state of the art in topic modelling techniques to index 19,000 letters of correspondences between 16th and 17th century Dutch scientists.

Other (high-level) text analysis methods, such as frequency-based corpus analysis to compare e.g. work from different authors or to investigate other stylometry characteristics, are also popular in the

[56] HiTime Project. http://ilk.uvt.nl/hitime/

[57] "Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body". https://en.wikipedia.org/wiki/Topic_model

e-humanities domain[58]. These methods are not domain dependent and fit more easily into the e-humanities researcher (search-based) toolbox.

The spectrum of tools to extract knowledge from unstructured historical data is wide. Important contributions are essentially domain-independent (Augenstein, Padó and Rudolph 2012), thus not particularly focused on historical text processing. Gangemi (2013) presents a recent and complete comparison of generic knowledge extraction tools for the Semantic Web, which will aid historical researchers working in the phases 2 (enrichment) and 3 (editing) of the historical information life cycle.

5.1.4 SEARCH AND RETRIEVAL

In *search and retrieval* we include systems that exploit semantic formalisms as a new way of indexing, querying and accessing historical data, instead of relying on the traditional text-based or keyword-based algorithms. This task matches the phase 4 (retrieval) of the life cycle of historical information (Figure 6).

It is not a coincidence that a high number of contributions that aim at extraction of structured entities from historical data also point at some desired system able to improve search and retrieval of such entities. Indeed, by means of constructing a semantic graph of historical knowledge, search and retrieval of that knowledge, as well as indexing systems that give exact pointers to the source in which particular historical entities are mentioned,

[58] The eHumanities Group. http://ehumanities.nl/

can be easily built and improved. The Agora (museum collections), BRIDGE[59] (historical TV metadata), CHoral[60] (historical audio metadata), Historical Timeline Mining and Extraction (HiTiME) (biographical events), Verrijkt Koninkrijk[61] (Dutch post-war social clusters concepts) and FDR Pearl Harbor (historical events around Pearl Harbor attack on 1941) projects are all good examples of the tendency to improve search and retrieval. Once the knowledge is successfully extracted from the historical sources and formalized appropriately, entities structured this way can be used for a graph-based search and retrieval (for instance through SPARQL queries, see section 5.3), although most systems use specific access methods (Ide and Woolner 2004). Other projects, like the H-BOT[62] project, use a natural language interface instead of a query system for retrieval of such historical structured knowledge.

Indexing of historical contents is another way of improving search and retrieval of historical data. Indexing and historical data storage systems have a long tradition in historical research (Boonstra, Breure and Doorn 2004). CLIO (Thaller 1980) is a traditional example of such a system, nowadays indexing is performed by XML annotation-oriented approaches, such as described by Robertson (Robertson 2009). These initiatives should consider the emerging RDFa, microformats and microdata technologies (see section 5.3) to study the ways they fit in the vast domain of historical text annotation systems.

[59] BRIDGE Project. http://ilps.science.uva.nl/ node/735
[60] CHORAL Project. http://hmi.ewi.utwente.nl/ choral/
[61] CLARIN-VK Project. http://verrijktkoninkrijk.nl/
[62] H-BOT Project. http://chnm.gmu.edu/tools/ h-bot/

5.2 INTEGRATION OF HISTORICAL SOURCES

In this section we analyze to what extent the contributions we presented consider the problem of data integration and use the Semantic Web to deal with it. The specific problems encountered are data model mismatching, schema incompatibilities and disparate source formats. Semantic interoperability has much to do with data integration, namely, how to commonly query and uniformly represent data that come from multiple sources (i.e. fitting several, probably non-compatible data models).

Heterogeneity of historical sources is especially present in social history projects. The North Atlantic population project (NAPP[63]), also deals with harmonization challenges, in which heterogeneity of sources requires intensive work on resolving data model inconsistency between datasets. In CEDAR the transcribed census tables all share the problem of structural heterogeneity.

The source material for the Historical Sample of the Netherlands (HSN[64]) database consists mainly of the certificates of birth, marriage and death, and of the population registers. From those sources the life courses of about 85.000 people born in the Netherlands during the period 1812-1922 have been reconstructed. Stored in a database and downloadable as files, this information forms a unique tool for research in Dutch history and in the fields of sociology and demography. As in the case of the HSN this type of sources is usually stored in archives, and, for the majority from a more remote past, not yet machine readable and

[63] North Atlantic Population Project. http://www. nappdata.org/napp/
[64] Historic Sample of the Netherlands (HSN). http://www. iisg.nl/hsn/

not easy to analyze with NLP techniques. There is one major pitfall in linking this kind of data: extracting data about persons, events, institutions, locations is one thing, but *linking* to their different instantiations (for instance different name spellings, or persons with the same name) and providing and keeping good documentation is the real challenge (Mandemakers and Dillon 2004).

Our (CEDAR[65]) project, located in the crossroads of the Semantic Web, statistical analysis and social history, exposes the Dutch historical census data in the Semantic Web. Censuses are a great source of information, but they present complex problems making harmonization of the data a cumbersome process. The work developed by Sieber, Wellen and Jin (2011) provides a deep analysis of how semantic heterogeneity can be addressed exclusively with semantic technologies, and describes how to achieve success in environments with very disparate data models. In the history-related domain of geographic information systems (GIS), already discussed in section historical-ontologies, Manso and Wachowicz (2009) provide an extensive review on current issues in interoperability.

5.2.1 CLASSIFICATION SYSTEMS

Multiple publications in classification systems (Esteve and Sobek 2003; Goeken, Bryer and Lucas 1999; Meyer and Osborne 2005) are especially aimed at solving interoperability and heterogeneity problems in historical data. Classification systems provide a

[65] CEDAR Project. http://www.cedar-project.nl/

standard mechanism to compare such data, but their specific implementation and effectiveness depends on the orientation towards source or goals of the historical data created in phase 1 of the historical data life cycle (Figure 4.2). When dealing with vast amounts of historical data, classification systems are a necessity in order to organize and make sense of the data. This entails an allocation of classes which are created according to certain relations or similarities. The main issue with historical classification systems in original sources is that they are not consistent over time, making comparative historical studies problematic. Historical census data is a typical example of this problem (Meroño-Peñuela et al. 2012). Major changes in the classification and coding of the different censuses, have hindered comparative historical research in both past and present efforts (Putte and Miles 2005). In order to deal with the changing classifications and vast differences at both national and international level, we need to connect the gaps between the datasets and conform to certain standard classification systems.

Currently several significant efforts have been made in this direction. The Integrated Public Use Microdata Series (IPUMS[66]) project for example faces the problem of bridging 8 different occupational classification systems and a total of 3200 different categories, containing the richest source of quantitative information on the American population. The NAPP project of IPUMS provides a machine-readable database of nine censuses from several countries. The main focus of the NAPP project is to harmonize these data sets and link individuals across different censuses for longitudinal and comparative analysis. Their linking strategy involves the use of variables which are supposed to not

[66] Integrated Public Use Microdata Series (IPUMS). http: //www.ipums.org/

change over time such as sex, birth year, name etc. In this process records are only checked if there is an exact match for some variables, such as race and state of birth. Other variables like age and name variables are permitted to have some variations. Another significant historical classification system is the Historical International Standard Classification of Occupations (HISCO). As occupations are one of the most problematic (due to differentiation, specialization etc.) variables in historical research, HISCO aims to overcome the problem of changing occupational terminologies over time and space. It encodes historical occupations gathered from different historical sources coming from different time periods, countries and languages, and classifies tens of thousands of occupational titles, linking these to short descriptions and images.

5.2.2 TRANSVERSAL APPROACHES

Finally, there are few but key contributions we have categorized as being *transversal*, because they cover a wide spectrum of the list of overlapping tasks between the Semantic Web and historical research. They also influence almost every phase in the historical information life cycle (Figure 4.1).

The CLIO system (Thaller 1980), a databank oriented system for historians, is the first of such contributions. CLIO was, for decades, *the* system for creating, enriching, organizing and retrieving historical knowledge from historical data in the pre-Web era. Although not using Semantic Web technologies, it had a strong emphasis on *semantics* as key for *structuring historical knowledge*.

In the Linked Data universe, the Agora project is one of such transversal contributions. It generates historical RDF of events extracted using NLP techniques from unstructured texts, uses it for enhanced search and retrieval, improves semantic heterogeneity and gives context by linking to other datasets. Similarly, the Verrijkt Koninkrijk (VK) and Multilingual Access to Large Spoken Archives (NSF-ITR/MALACH[67]) projects perform these tasks in their particular domains. The FDR Pearl Harbor project also contributes on this line, but additionally opening the very promising field of historical knowledge inference through the formalization and usage of historical OWL ontologies. All these are good examples on how historical data get

[67] MALACH http://malach.umiacs.umd.edu/

much richer when their semantics are explicitly expressed and they are interlinked through standard vocabularies and ontologies.

Regarding tools, the Armadillo architecture of Semantic Web Services and the Fawcett toolkit contain the generic plot behind all these contributions, and cover the whole pipeline of semantic historical data management. The latter extracts RDF event-oriented triples from unstructured texts, and additionally allows historians to install a full semantic toolbox with widgets to experiment with their data. Open Refine, in combination with its RDF-export plugin, allows the extraction, transformation, modelling and publishing of (historical) Linked Data when the sources come in tabular format.

Additionally, the theoretical study by Boonstra, Breure and Doorn (2004) envisages possibilities on how the Semantic Web can enhance research by historians. It constitutes, besides, a major work on the evolution of historical computing, e-history and historical information science, and gives a deep intuition on how computer science can help to solve ancient problems in historical research.

5.3 SOLVING HISTORICAL PROBLEMS - A REFLECTION

In this section we point to the open historical data problems revisited in this chapter which are addressed or solved by the Semantic Web contributions reviewed in our survey. The mapping between the open problems and the tasks is shown in Table 5.1, see the following page.

Open historical data problems	Historical Ontologies	Linking historical data	Text processing and mining	Search & retrieval	Classifica-tion systems	Transversal approaches
Historical Sources	X	O				
Relationship between sources		X			X	
Historical analysis						O
Presentation of sources						O

Table 5.1 Mapping between the open problems of historical data and the surveyed contributions in historical Semantic Web. The sign X indicates that the problem is directly addressed in the Semantic Web task. The sign O indicates that the problem is indirectly or partially addressed in the Semantic Web task.

The first interesting result is that some of the problems identified in *historical sources* are mostly solved by the approaches we review in historical ontologies. Concretely, our perception is that the structuring of historical data and the development of historical data models have been a success due to the creation of *standard* vocabularies and classification systems (a vocabulary with hierarchy and structure). These *types* of ontologies aid historians to describe, at least, the baseline historical entities and relations in historical domains. The large number of projects exposing historical Linked Data on the Web using these ontologies prove their usefulness and success. There is space, though, for improvement. Although it is commonly agreed that current historical ontologies model the core semantics of historical research, authors also agree that they are still *scarce* and need further development (Ide and Woolner 2007; Owens et al. 2009; Ashkpour, Meroño-Peñuela and Mandemakers 2015).

As part of the problems in historical sources (see section 4.5.1), provision of historical background knowledge has been successful only partially. The Semantic Web infrastructure (Linked Data cloud, SPARQL endpoints on historical data) is set up and running. But the amount of historical data available is still too small to give good support to any historian creating historical datasets in the beginning of the life cycle. Consequently, little background knowledge can help these historians in solving e.g. errors or inconsistencies at that phase. Similarly, the generic infrastructure for provenance, publishing and retrieval in the Semantic Web is very mature and extensively used in other domains, but scarce or non-existing in the historical domain although being identified as a very important requirement. In our

harmonization approach this aspect is given specific attention (chapter 7). The provision of provenance (e.g. accountability) on historical datasets needs to be guaranteed in projects using semantic technologies to publish historical data (a key aspect in historical research).

Solutions to the problem of *relationships between sources* (see section 4.5.2) are probably the greatest achievement of the application of semantic technologies to historical research. The large number of projects linking historical data we survey proves that the Semantic Web delivers working solutions to the problem of connecting isolated historical data sources. The usage of developed ontologies and vocabularies has been key to this end. Additionally, the existence of classification systems helps on data comparability in the Semantic Web. Because we see that the body of historical knowledge in the Semantic Web is still small, we expect the problem of finding related links between historical entities and datasets to grow in the future, although the Semantic Web has generic solutions for this (Shvaiko and Euzenat 2013).

The problems in *historical analysis* (see 4.5.3) and *presentation of sources* (see 4.5.4) are only partially addressed in approaches we have classified as transversal. These works cover a wide spectrum of the life cycle of historical data, including analysis and presentation (phases 5 and 6). Consequently, they deal with some analyses and visualizations. However, there is a lack of contributions tackling directly the problem, or considering explicitly historical research requirements with respect to analysis and visualization. The transversal tools are hence very generic, and they could be considered inappropriate for some historians.

Therefore, it is very important to distinguish what analysis requirements are specific to historical research, and which ones are domain-independent. Our hypothesis is that these problems overlap only partially with the goals of the Semantic Web (i.e. representing and linking meaning on the Web). However, historians could benefit from analysis and visualization tools for historical semantic data, not as specific as project-oriented, but not as generic as domain-independent.

In Table 5.1 we see that all identified problems have Semantic Web tasks attached to them, and provide some solutions. However, not all tasks are mapped to some historical *open* problems. Concretely, the tasks of text processing and mining (section 5.1.3) and search and retrieval (section 5.1.4) do not seem to solve any of the identified problems. Why do we find contributions on these areas? First, although not being identified by historians as primary problems, they constitute secondary problems that need to be solved when representing and linking semantic historical data. These problems are not exclusively historical, but they needed to be re-implemented in the Semantic Web realm (e.g. natural language processing for extracting historical RDF triples, SPARQL to query historical semantic data on the Web). Secondly, the goals these tasks aim at were quite well solved in historical research before the inception of semantic technologies (e.g. manual input of historical data, SQL queries in historical relational databases), and thus historians did not consider them into the primary problem space.

5.4 OPEN (INTEGRATION) CHALLENGES

The use of semantic technologies has contributed significantly to solving the open problems of historical data (4.5). However, there is a lot of room for improvement. The open problems are being addressed, but they are far from being solved until they get additional attention. The scarce amount of historical data on the Semantic Web is a good example. Other problems, some more specific, some more generic, could be also tackled with semantic solutions. In this section we explore some aspects of the Semantic Web that have not been used yet or could be furtherly exploited in historical research.

SEMANTICS OF TIME, CHANGE, LANGUAGE, UNCERTAINTY AND INTERPRETATION

Classifications and ontologies in history do exist, but not for all areas, not in Semantic Web languages and not always agreed upon. Although several historical ontologies have been developed (see Section 5.1.1) these models are insufficient for the vast amount and variety of historical data that still has to be published in the Semantic Web, especially when key issues for historians like *interpretations* or *evidences* need to be modelled and conveniently linked. Historical ontologies and vocabularies have been a reality in recent approaches. Ontologies describing classes and properties of some historical concern, such as concepts around the Pearl Harbor attack in 1941 (Ide and Woolner 2004), are an exciting modelling exercise for researchers but also a necessary step for better structuring historical information in the Web. Ontologies offer a way of controlling e.g. the classes, properties and terms that

the community uses as a standard for describing factual and terminological knowledge about history. Designing good ontologies for historical domains is also an area with plenty of challenges: how can ontologies comprise the many conceptions of history depending on the temporal dimension of events described (Ide and Woolner 2007)? Moreover, how can differences in meaning and relations between concepts be traced, as time and historical realities change these concepts (Wang, Schlobach and Klein 2011)? To what extent do these differences relate to the complexity of the language (e.g. Latin, Middle languages) and uncertainty (e.g. fuzzy dates and locations)? These questions, which comprise semantic technologies, knowledge acquisition and knowledge modelling techniques, are not yet completely understood and are a significant challenge in semantic research. On the other side, over the centuries, different dictionaries, thesauri, classification systems have been developed. How to mount those specifically grown ordering principles to the Web in a way that makes them explorable and linkable to other ontologies is an interesting challenge which requires a close collaboration between historians, knowing and designing those specific tools, and computer scientists, often relying on much broader and generic ontologies.

REASONING

From the point of view of *Linked Data*, ontologies are designed in order to control the terms in which datasets may express data, as well as the data model in which these data are represented. However, in a more Semantic Web perspective, one may expect these ontologies and vocabularies to facilitate new knowledge

discovery; that is, to make explicit some implicit fact that was not trivial to deduce for the human eye, especially in big knowledge bases.

Reasoning is one of the key mechanisms of the Semantic Web still to be used in historical research. The absence of specific methods and tools for automatic historical inference, so that new, *implicit historical knowledge* can be derived, is another issue. We claim that reasoning could be fundamental for historical analysis (4.5.3) and tasks in the phase 5 (analysis) of the historical information life cycle (Figure 4.2).

Historical ontologies can be used to facilitate historical knowledge discovery using reasoners (i.e. software that can infer logical consequences from a set of asserted facts). Assuming that a particular domain is completely formalized as historical ontologies, then it is possible to run a reasoner on these ontologies to produce derived, implicit rules and facts that were not present in the original model as explicit knowledge (i.e. specifically encoded in the ontology). For instance, if an ontology describes, on the one hand, the fact that a letter was sent from one government to another, and on the other hand, the fact that governments have a person responsible of sending and receiving letters, then it may be possible for the reasoner to infer what concrete persons sent and received what letters. As the knowledge base grows, implicit knowledge is not evident anymore and reasoners can facilitate an enormous work and produce high-value pieces of historical knowledge in ways which are currently not explored enough.

Since historians have different interpretations and not always a clear research question when starting an investigation, abductive reasoning (i.e. given the conclusions and a rule, try to select possible premises that support the conclusion) may be more convenient than deductive reasoning (i.e. deduce true conclusions given a premise and a rule) in historical research (Charniak and McDermott 1986; Hobbs et. al 1993). These would revert the order of some phases of the life cycle of historical information, generating a more bottom-up, data-based generation of hypotheses supporting evidence. The impact of abductive reasoning in historical research and its relationship with the life cycle needs further study and clarification. The introduction of any kind of reasoning in the life cycle needs to be done with the goal of supporting, not replacing, the task of the historian, who must keep control of the implementation of the different phases.

LINKING MORE HISTORICAL DATA

We show in section 5.1.2 (linking historical data) that great efforts are being devoted to publish historical Linked Data. However, the amount of structured historical knowledge available on the Web is still insufficient to aid tasks that need high amounts and different kinds of historical background knowledge. While many different data and information sources exist, they are not always interlinked. This isolation of historical data sources obstructs that they can be found, but it also constrains how they can be further processed and connected.

One of the big claims of linked data is that, by linking datasets, relations established between nodes of these datasets highly enrich

the information contained in them. That way, browsing datasets is not an isolated task anymore: by allowing users (and machines) to explore entities through their predicate links, data get new meanings, uncountable contexts and useful perspectives for historians.

For example, consider a scenario with three different SPARQL endpoints exposing RDF triples of a census with occupational data, a historical register of labour strikes, and a generic classification system for occupations (in the context of one particular country, for instance). Suppose that: the occupational census of the data exposes triples with countings on occupations (for example, how many men and women worked in a particular occupation in a concrete city), the historical register of labour strikes contains countings on how many people participated in labour strikes (number of women and men, per occupation and city), and the generic classification system harmonizes the occupational titles between both previous datasets (for example, gives a common number for representing occupational titles that may vary between census occupations and labour strike occupations). Then, it is clear that several SPARQL queries can be constructed to give very meaningful and interesting linked data to the historian. For instance, such a query may return, given a city and an occupation code, which ratio of men and women followed a particular well-known labour strike. Another SPARQL query may return an ordered list of historical labour strikes by relevance, according to several indicators (strike successfulness ratio, total number of workers on strike, density of people on strike depending on the location, etc.). It is obvious that the possibilities increase if we think of more related historical sources

to link, like datasets describing historical weather or historical geographical names and areas.

FLEXIBILITY OF DATA MODELS

It is considered to be a bad practice in historical research not doing the historical data modelling at phase 1 of the historical information life cycle. Data models are used to provide an abstract data structure that organizes elements of data and standardizes how they relate to one another (e.g. a database models used in relational databases). The choice of a particular data model to represent historical data is a critical issue for most historical computing projects. The election of some appropriate data model may seem a good design decision at some stage of the project. However, new requirements, research directions or stakeholder priorities may convert that data model into an obstacle more than an aid. Accordingly, flexibility of the model used to represent historical domains is desired to avoid restructuring entire databases. Comparison in historical research requires flexibility of the models to be able to match them to one another or allow different interpretations. At the end, that enforces historians to make their data selection and processing dependent of a certain data model that cannot be easily replaced or altered if needed. This happens usually in environments with changing and creep[68] requirements (Jones 1996).

Applying semantic technologies and Linked Data principles to historical data may have a major advantage regarding historical

[68] https://en.wikipedia.org/wiki/Scope_creep

data models, providing flexibility at the historical data modelling phase. Two different approaches regarding historical data modelling have been followed traditionally in historical computing: the *source-oriented*, and the *goal-oriented* representation (described in chapter 3). In this section we aim to identify whether semantic technologies allow a more flexible representation of historical data? RDF is known for storing the 'middleware representation' of alternative views on the data. These views can be modelled according to any particular historical interpretation as needed. This way, the decision of what data model suits the historical source better can be postponed until the very end of the life cycle, or adopted as early as necessary.

Moreover, additional questions arise when considering the traditional perspectives on data modelling, i.e. the conceptual, logical and physical data models. These perspectives help in detaching data management technology, like relational databases or RDF triplestores, from conceptual schemas (i.e. the semantics of a domain). While conceptual data models are currently shared on the Web as e.g. historical ontologies, the flexibility of the whole modelling stack towards semantic changes needs to be better understood.

Most of the time, historical data are modified in the life cycle of historical information processing. If update, enrichment, analytic and interpretative operations are not controlled (e.g. standardized), these transformations lead to different historical data representations which can hardly be related to each other anymore, nor in terms of provenance nor in terms of relatedness. Another issue is supporting data transformations under two

constraints: (a) without modifying source data (so the originals stay intact); and (b) with tracking of changes. Consequently, destructive updates are a major concern when selecting, aggregating and modifying historical data. On the one hand, modifications to specific encodings in formats such as CSV, spreadsheets and XML, do not support non-destructive updates, and version control systems are necessary to retrieve previous states. On the other hand, relational databases can be inefficient when querying all recorded transformations, edits and manipulations. Non-destructive updates are well supported by current Semantic Web technology like SPARQL, the query language of RDF. SPARQL allows the construction of RDF triples according to the supplied graph pattern, facilitating data transformations without altering consistency of previous states in the knowledge base.

5.5 CONCLUSION AND LESSONS LEARNED

In this chapter we present a general overview of semantic technologies applied to historical research. We describe a general approach to historical research and the Semantic Web, and motivate why the combination of the two is an interesting field of research. We introduce core elements of historical research, such as the life cycle of historical information, several classifications for historical data, and the open problems shared by historians and computer scientists. Then, we overview contributions to the young historical Semantic Web in form of papers, projects and tools, articulating the work into several tasks and trends. We provide a mapping to see to what extent the work on these tasks

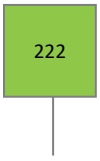
is helping to solve the open problems of historical data and historical research. Finally, we dig out a list of interesting open challenges for the future, like working out the semantics of critical aspects for historians, such as interpretation and time, and encouraging reasoning in the historical Semantic Web.

It is interesting to observe the sparsity in tables A.1, A.2, A.3 and A.4 (see Appendix). There is a significant difference in the number of empty spaces (i.e. specificity of the contributions) between tables A.1 and A.4 (papers and tools, ontologies), and tables A.2 and A.3 (projects and online resources). While the former set has essentially lots of holes, the latter has lots of complete lines. The reason for this is probably the specificity researchers think research papers and useful tools need. Usually written by computer scientists, papers and tools need to be grounded and tackle a very concrete problem to be worth written or implemented. On the other hand, projects (Table A.2) are written in a very generic way covering all tasks, with probably intensive participation of historians and clear aims to solve the whole pipeline. In practice, though, these goals are materialized in very concrete research contributions. This leads us to think that Semantic Web solutions need very specific requirements in order to be correctly deployed in history. They need to be applied to historical data in a complex, layered and properly adapted pipeline. Good practices and standards, and their relationship with the life cycle of historical information, are still needed for the field to continue evolving.

We show how the Semantic Web and history communities understand the need for representing inner semantics implicitly contained in historical sources, and how these semantics can be

conveniently identified, formalized and linked. With the appropriate pipelines, algorithms can extract entities from digital historical sources and transform these occurrences into RDF databases, according to some historical ontology or vocabulary. These entities can be linked between them and with other historical Linked Data, contributing to an open, world-wide, online persistent graph of historical knowledge: an historical Semantic Web. The work presented in this survey contributes in one phase or another in this graph-building pipeline.

The challenge of the realization of a historical Semantic Web meeting as many requirements as possible may bring new facilities for a number of stakeholders. On the one hand, humanities researchers, also outside history, will be able to integrate the historical Semantic Web to their own information life cycle. They will be able to search, retrieve and compare historical knowledge and use it for the construction of their narratives, still the final outcome of historical research. On the other hand, practitioners will be able to search new data sources thanks to the open character of the Semantic Web, and develop (history-aware) applications for public institutions, private companies and citizens.


222

Part III

HARMONIZATION OF HISTORICAL CENSUSES USING LINKED DATA

6. HISTORICAL CENSUS DATA HARMONIZATION AND THE SEMANTIC WEB

After describing the various uses of Semantic Web technologies in historical research we now focus on its application and benefits for the harmonization of socio-economic historical data such as the census. In chapter 2 we have described the main challenges of moving towards a harmonized census database, building on aggregate data. In this chapter we describe *how* we aim to work towards a harmonized dataset for the censuses of 1795-1971. We propose a specific approach and model in creating an interlinked census dataset in the Semantic Web using the technology of the so-called Resource Description Framework technology (RDF).

In this chapter we elaborate on our method of using RDF to model, expose and harmonize the aggregate data from the historical censuses over time. We explain our motivation for using RDF as the modeling technique and how we aim to harmonize the aggregate census data. We next present the three tier data model we have created to allow a flexible and source-oriented harmonization approach when dealing with historical data. After this we describe the conversion process from source data (the Excel tables) to RDF. Once converted into RDF we show some preliminary uses and advantages of having everything in one system (as opposed to the unconnected Excel files).

This chapter is based on work published in (1) Ashkpour, A., Meroño-Peñuela, A., Mandemakers, K. The Aggregate Dutch Historical Censuses: Harmonization and RDF. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 48(4), pp. 230-245. (2015). As the main author I was responsible for the creation of the harmonization model, extensive usage and testing of our approach, tools and harmonization solutions presented. Next (2) in chapter 6 this article is enriched with original work in all sub sections and expanded with new text in 6.3.4 and 6.4.

6.1 HARMONIZING HISTORICAL CENSUS DATA IN RDF

The "Semantic" Web aims to enable the sharing of content from databases and other structured data sources which as such are not directly published on the Web. As noted earlier, the Semantic Web is the collaborative movement and the set of standards that pursue the realization of this vision. The Resource Description Framework is the basic layer on which the Semantic Web is built. RDF is considered 'as the standard model for data interchange on the Web and has features that facilitate data merging, specifically supporting *the evolution of schemas over time*[69], meeting the harmonization requirements of census data.

Using RDF as *the* knowledge representation model in our harmonization approach, we have developed tools and methods to create a *one to one* copy of the structure and contents of the original Excel files in the form of a (graph) database and to separate the harmonization process from the data itself (Mandemakers and Dillon 2004, Meroño-Peñuela et al. 2012). We facilitate the different views / interpretations on our data by creating a three-tier architecture in RDF. In this architecture we make a strict distinction between the original data, the annotations and harmonizations. Doing so, we also guarantee provenance and access to the original data in a source-oriented approach which has always been a point of attention in the digitization and transcription of the Dutch historical censuses.

There are several reasons why RDF is chosen as the data system in which we model, publish and query the Dutch census dataset.

[69] http://www.w3.org/RDF/

First, a graph data model like RDF is appropriate when the dataset suffers from structural heterogeneity. This is especially true in our case, where data spans two centuries and the schemas behind the tables changed substantially from one census to the other. In fact, we have 2,249 unconnected tables with different hierarchical structures which we aim to preserve. Moreover, there is no RDF requirement corresponding to SQL's structural constraint that every row in a relational database Table is defined in the same schema. Therefore these 2,249 census tables can be represented with diverse RDF graphs that match their diverse structure. Without constraints on meeting an overall agreed schema makes it possible to follow a truly source-oriented approach. This is especially useful to extend and particularize descriptions of resources. For instance, variables or the census tables can be more concretely defined with the specificities that might apply at different points in time. Second, the RDF model allows data publishers to easily link their datasets to other RDF datasets, since RDF and SPARQL (the RDF query language) were designed to merge disparate sets of data on the Web. For example, the SPARQL query in Figure 6.1 illustrates how the linkage between the Dutch historical censuses in RDF and other sources of Linked Data on the Web are used to extend information on Dutch municipalities. The example shows how we compare the 1889 and current populations of the municipalities, using the CEDAR and DBPedia's[70] Linked Data endpoints (i.e. RDF Databases):

[70] http://dbpedia.org/ a Linked Data equivalent of Wikipedia

```
prefix qb: <http://purl.org/linked-data/cube#>
prefix cedar: <http://lod.cedar-project.nl:8888/cedar/resource/>
prefix maritalstatus: <http://bit.ly/cedar-maritalstatus#>
prefix sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#>
prefix sdmx-code: <http://purl.org/linked-data/sdmx/2009/code#>
prefix cedarterms: <http://bit.ly/cedar#>
prefix dbpprop: <http://dbpedia.org/property/>
prefix owl: <http://www.w3.org/2002/07/owl#>

SELECT ?municipality (SUM (?pop) AS ?oldpopulation) ?currentpopulation
WHERE {
SERVICE <http://lod.cedar-project.nl/cedar/sparql> {
  ?obs a qb:Observation.
  ?obs cedarterms:population ?pop.
  ?obs sdmx-dimension:refArea ?municipality .
  ?slice sdmx-dimension:refPeriod "1899"^^xsd:integer .
  ?slice cedarterms:censusType "VT" .
SERVICE <http://www.gemeentegeschiedenis.nl/sparql> {
  ?municipality owl:sameAs ?currentmunicipality .
}
SERVICE <http://dbpedia.org/sparql> {
  ?currentmunicipality dbpprop:population ?currentpopulation .
}
}
```

Figure 6.1 - Example SPARQL query using two different sources (i.e. datasets) to answer a question

By harmonizing the Dutch historical censuses in RDF we build a hub of socio-historical information, where census numbers and variables can be easily linked to historical classifications of occupations, municipalities, regions, labour strikes and religions, as well as other cross-domain datasets such as DBPedia. By creating such links to other datasets, extended and enriched census information can be retrieved. Combining data from the linked sources, we can for instance ask a question which is outside the scope of the census itself: i.e. "number of workers per occupation and year, versus the number of labour strikes per occupation, year and municipality or region".

We distinguish our harmonization approach using RDF in different ways. As we have shown in chapter 3, over the past years different efforts have been undertaken using RDF technology for greater census utilization. However, as we found, all these projects merely harmonized data *within the domain* of each census year, using *microdata* and *contemporary* censuses as a starting point, and have focused mainly on *publishing* the data. In the following, we will explain our harmonization approach and how we have used RDF in a novel way to *model* and *harmonize* aggregate historical data *over time and space*.

6.2 A THREE-TIER DATA MODEL

In order to deal with the challenges of the historical censuses, we model our dataset in a three-tier architecture according to the multi-tier architecture principles where layers are logically

separated. In our model the architecture consists of a *harmonization* layer, a *raw* data layer, and an *annotations* layer (see Figure 6.2). The dependencies between the layers are represented in Figure 6.2 with directional arrows. An arrow from A to B means that structure and data from A needs to be linked to structure and data from B.

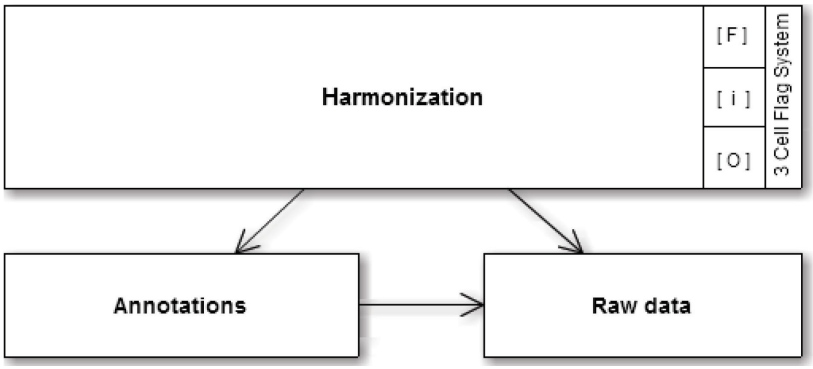


Figure 6.2 -Three-tier Harmonization Model

In our harmonization approach, we separate the data for several reasons. First, the census source data contained in the raw data layer should be preserved, even if it contains errors, in order to be able to provide data provenance in the RDF system and to have a digital copy of the source. Second, as mentioned before, the process of correcting the census data is an ongoing process and will continue to do so in the future. Accordingly, we have designed a system which is flexible enough to feed new annotations into our three layered model. We also aim to allow suggestions for changes to the data via an online interface. To control the quality of the data, the annotation system is designed in such a way that the suggested changes are only accepted after manual review. In order

to cope with the different type of annotations in our dataset, we have extracted, standardized and modeled the annotations according to a RDF annotation standard. How we process these corrections / annotations will be elaborated in the next section. Finally, harmonization is a dynamic process that affects how raw data are interpreted, transformed and presented, and which may need to be customized according to multiple research requirements. By storing the different harmonization practices in a separate layer: allows us to modify the harmonization procedures as we go in an iterative manner, without affecting the underlying raw data. Moreover, due to the ambiguity related with aggregate census data harmonization, this approach allows us to provide several solutions to each particular problem.

6.2.1 RAW DATA LAYER

The raw data layer consists of a *one to one* copy of the original Excel sheets (see Figure 2.2). This means that we present the original source data in a strict source-oriented approach. The 2,249 Excel tables with their different structures and layouts are stored in the *raw data layer* in the form of RDF graphs. Since the data contained in a census Table is statistical data, we have designed a data model around the central concept of the Table cell (i.e. a data cell in our Excel tables), according to the W3C RDF Data Cube vocabulary[71]. RDF Data Cube is *the* standard for modeling and publishing multidimensional data, such as statistics, in the Semantic Web. Accordingly, it provides means to "publish multi-dimensional data, such as statistics, on the web in such a

[71] http://www.w3.org/TR/vocab-data-cube/

way that they can be linked to related datasets and concepts[72]" (Cyganiak et al. 2014).

Although the *layout* and *structure* of the Excel tables differ significantly across our dataset they contain the same basic structure of three areas, namely: cells containing the *data* as such and, *column* and *row headers* defining the data. Although humans can easily spot where the numbers and variables are, we need to specify for each Table where the *columns*, *rows* and *content* areas start and end. This is done by way of so-called *bounding boxes* which we use to define the Table layout of the raw data layer. The use of bounding boxes helps us to keep track of the different Table structures and deal with structural heterogeneity across our dataset. Exploiting this *shared characteristic* of the tables allows us to apply the same approach in converting all Excel tables to RDF. In section 6.3 we go into the details of creating these raw representations in RDF.

6.2.2 HARMONIZATION LAYER

In our harmonization data model we separate the harmonization layer from the original data at all times. By doings so we are able to harmonize the data in an *iterative* way and gradually explore the peculiarities of data which has undergone so many changes. Following this approach we are able to provide on a cell level the provenance of our harmonizations and always trace back our decisions to the original sources.

Harmonizing aggregate data such as the Dutch historical censuses

[72] https://www.w3.org/TR/vocab-data-cube/

draws upon on a *combination* of different harmonization practices including resolving inconsistencies, data cleaning and correcting, restructuring of the data but also adding redundancy to the Excel tables to make values or variables explicit. Next to these types of harmonization practices we apply a combination of bottom up and top down approaches in order to further harmonize the census and make consistent classifications and variables. As discussed earlier, these include the creation of standard vocabularies, the construction of variables across the different censuses, creating new variables and values and connecting them to existing classification systems. We store all these types of created data (mutations, standardizations, variable creations, classifications etc.) in the harmonization RDF layer which can be enriched and modified as a continuous iterative process without compromising the underlying data. In the next chapter (7) we zoom in and take a closer look on this harmonization layer by describing the different steps and practices which (source-oriented) harmonization actually entails.

Thus far we have described harmonization as the process of creating a unified and consistent data series from various census tables. This process of creating requires interpretations (changes) to the data. In order to deal with these interpretations we have a data layer which we call the *Three Cell Flag System*. This is the nucleus of our approach and means that we have three variables for each cell-value of the census, namely; the original value, the interpretation (which may be the original or a new value) and a flag, indicating the nature of the interpretation. For example, if a cell has the original value of '39', and cross validation showed that '39' was a typo and should be '93', the interpretation gets the

correct value '93' and the flag will indicate that the corrected value was based on cross validation over the row and column totals. In other cases in which we accept the original value as correct, resulting in the same values for the interpretation and the original, we indicate this fact in a flag (we elaborate and build on this principal using our GapFiller example in section 7.2.6).

When we combine two variables to create a higher level aggregation for harmonization purposes, we in fact create a new layer. Building on the example illustrated in Figure 2.4, where we harmonized the 'stonecutters' and 'diamond workers' of 1899 into one group to make it comparable with the census of 1889, we combine the values of both groups of 1899 into one and the same value both for the original value and the interpretation. The interpreted value may be further modified in this action, indicated by a different flag value. The other way round: splitting a value of one group into values for two subgroups is different in that we immediately interpolate to achieve two interpreted values, where the flag indicates the rule on the basis of which we have split the original value. By introducing a source-oriented harmonization approach we aim to tackle these problems and create a *flexible* system which makes such different interpretations possible while keeping full provenance of our actions. In chapter 7 we go into the details of such an approach and present a source-oriented harmonization *workflow*.

6.2.3 ANNOTATIONS LAYER

Throughout their lifespan the censuses have been annotated in different ways, applying no consistent system, logic or provenance

to how and why the annotations were made. Scattered and even sometimes hidden in different tables, we encounter annotations which were source made (e.g. annotations printed in the original books to note that females were included with the male population instead of having the usual separate column for females), made during data entry (e.g. annotating that some specific Figure could be wrong) and corrections made after data entry (e.g. correcting probable mistakes based on existing annotations or newfound problems). Moreover, the way these different types of corrections were implemented in the Table conversion to Excel differs greatly across the tables and census years. We find annotations which were made as cell comments in Excel, as notes or even placed in a separate Excel sheet with a reference to the changed value in a cell. Because of this lack of structure and predictability, we cannot handle annotations as raw data. Instead, as a preliminary step, we extract all annotations from the Excel tables, standardize and model them in the annotation layer of our three tier model, using the W3C Community Open Annotation Core Data Model standard[73]. The created annotation layer is then *linked* with the raw data layer (see Figure 6.2). For provenance purposes we also attach an author and other information to each annotation. We flag the contents of the annotation to indicate the specific issue of this annotation using a second system of the aforementioned Flag System.

Table 6.1 gives an illustration of the flagging of the most common annotations. Information contained within these annotations can be used to make interpretations in the harmonization layer and will be flagged with a content that refers to the used annotation.

[73] http://www.openannotation.org/spec/core/

Flag	Description
1	Incorrect number
2	Source error – Sum does not add up
3	Source error – Name misspelled – Corrected
4	Source error – name misspelled
5	No value
6	Number includes – Sheds
7	Not Readable

Table 6.1 - Annotation classification based on a subset of the data

6.3 FROM ORIGINAL CENSUS TABLES TO LINKED DATA – CREATING HISTORICAL DATABASES IN RDF

In the following section we describe how we manually prepare the various unconnected (census) Excel tables for RDF conversion. The first stage in moving towards a harmonized database from the Excel sheets is the conversion of the tables into RDF, using a script called TabLinker[74] (Meroño-Peñuela et. al 2012). The data we produce in RDF is an exact representation of the underlying source data, maintaining the relations and hierarchies found in the original Excel tables.

6.3.1 SUPERVISED CONVERSION PROCESS

When moving towards an RDF database several alternative systems are available for researchers to use. The tool we use, i.e. TabLinker, converts the original structures of the Excel tables into an RDF graph for each census Table. To maintain the structures of the original tables, TabLinker needs to define different styles for each Table to link all *cell values* to the corresponding *columns* and *rows*. Using standard functionalities in Excel, we *color/style* the (bounding) boxes of our data manually by defining the columns, rows and cell areas of each Table (see Figure 6.3). Therefore, although the layout of the tables differ greatly they *do* share common characteristic (rows, columns, data areas) which we can use to style all tables uniformly into RDF.

[74] https://github.com/Data2Semantics/TabLinker

Figure 6.3 - Marked census Table with TabLinker (translated for illustration purposes).

This coloring/styling process is very straightforward and creates a faithful (one to one) representation of the tables in RDF. Our *TabLinker* defines the following styles:

- ***Title*** marks cells that contain the Table title and description, placed at the top left of the tables (this style is transparent and illustrated by a checkered version in Figure 6.3).

- ***Data*** marks the data cells with the actual census numbers (the white colored section in Figure 6.3). Since all measurements in the dataset are counts, these numbers are qualified as integers (xsd:integer) during the conversion; additional metadata is also attach to make explicit that these numbers represent *population counts,* using the property qb:measure provided by the RDF Data Cube vocabulary. Empty *data* cells are counted as zero's.

- ***ColHeader*** marks the column headers of the Table just above the content of the cells (the light blue colored section in Figure 10). These headers contain the values for different variables such as age ranges, residence status, marital status, sex, etc.

- ***RowHeader*** marks cells with row headers, usually placed at the left of the Table (the caramel colored section in Figure 6.3). These cells usually contain values for geographical variables like municipality.

- ***HRowHeader*** marks cells with hierarchical row headers (the purple colored section in Figure 6.3). This style is similar to row headers, with the difference that these cells form hierarchies or taxonomies (e.g. occupations of class *I*, subclass *a*, group *Diamond workers*).

- ***RowProperty*** marks cells with the names of the row variables, placed at the upper left of the Table (the dark blue colored section in Figure 6.3). These variables are usually not made explicit in the censuses. For example, the cell containing the string *Gemeente* (municipality in Dutch) is marked as a RowProperty, since it denotes the name of the variable (municipality) whose *values* are contained in ***RowHeaders*** or ***HRowHeaders*** in the cells below, like *Amsterdam* or *Haarlem*.

- ***Metadata*** are used to mark any additional defining data that the tables may contain, like references to column or page numbers of the census books (the orange colored section in Figure 6.3).

By applying these styles TabLinker first generates one 'tablink:DataCell' for each data cell (i.e. cells marked as Data in Figure 6.3), attaching its value (the actual population count / numbers) to the 'tablink:sheet' (i.e. Table in Excel) it belongs to. Secondly, the observation is linked with all its corresponding column and row headers (i.e. cells marked as RowHeader, HRowHeader, and ColHeader in Figure 6.3). Next to this, we create resources that describe the column and row headers, their

types, labels, cell positions in the spreadsheets and hierarchical parent/child relationships with other headers (which proves to be essential when inspecting the RDF generated data, see 'inspection stage' in section 7.2.2).

6.3.2 ALTERNATIVE SYSTEMS

Next to TabLinker, several alternative systems and methods have been made available by other projects in order to move data from spreadsheets towards an RDF database. We categorize such related work into two different types of contributions. These contributions can be divided into (a) workflows and tools for converting Excel/CSV data to RDF data and (b) methods for enriching these tabular-converted RDF graphs.

Currently there are many tools that convert tabular data such as the censuses to RDF[75] format. For tabular formats such as CSV and HTML tables, the data can be turned into RDF with dedicated tools (Lebo and McCusker 2012; Muñoz, Hogan and Mileo 2013). Larger frameworks, like Open Refine and DERI's RDF plugin (DERI 2015; Morris, Guidry and Martin 2015), Opencube (Kalampokis et al. 2014) and Grafter[76] cover even more tabular and structured data formats, such as Excel, JSON (for Javascript), XML, and Google Data documents. However, none of these are well suited for the conversion of historical tables. These types of tables are often presented in an "eccentric" layout with spanning and hierarchical headers, multidimensional views and arbitrary data locations, which does not match the regularity

[75] https://github.com/timrdf/csv2rdf4lod-automation/wiki
[76] http://grafter.org/

of modern tables. To the best of our knowledge, only TabLinker (Meroño-Peñuela et al. 2012) supports the characteristics of the census tables.

On enriching the RDF data coming out of statistical tables, Venetis et al. (2011) annotate Web tables using labels and relationships automatically extracted from the Web, to augment the semantics and improve access. In another example HTML tables are used to extend aggregated search results (Balakrishnan et al. 2015; Yakout et al. 2012) and to insert Web Table data into word processing software. Enriching RDF graphs with missing temporal information has also been given attention in publishing historical data as RDF. For instance, the authors of Hoffart et al. (2013) created a knowledge system that automatically integrates a spatio-temporal dimension from Wikipedia, GeoNames (a greographical database covering counries and place names) and WordNet data (a large lexical database of English). Similarly, Rula et al. (2014) proposes a generic approach for inserting temporal information to RDF data by gathering time intervals from the Web and knowledge bases. In another example, Fionda and Grasso (2014) focus on using the temporal aspect of Linked Data snapshots to keep track of the evolution of data over time.

6.3.3 GRAPH REPRESENTATIONS OF THE DATA

After converting the censuses to RDF data we create millions of *triples* according to the Resource Descriptions Framework model. These triples *together* form RDF Graphs. For each census Table we generate three RDF (large interconnected) graph-systems,

shown as three layers in Figure 6.2, i.e. raw, annotation and harmonization layer (where the latter is empty at this stage).

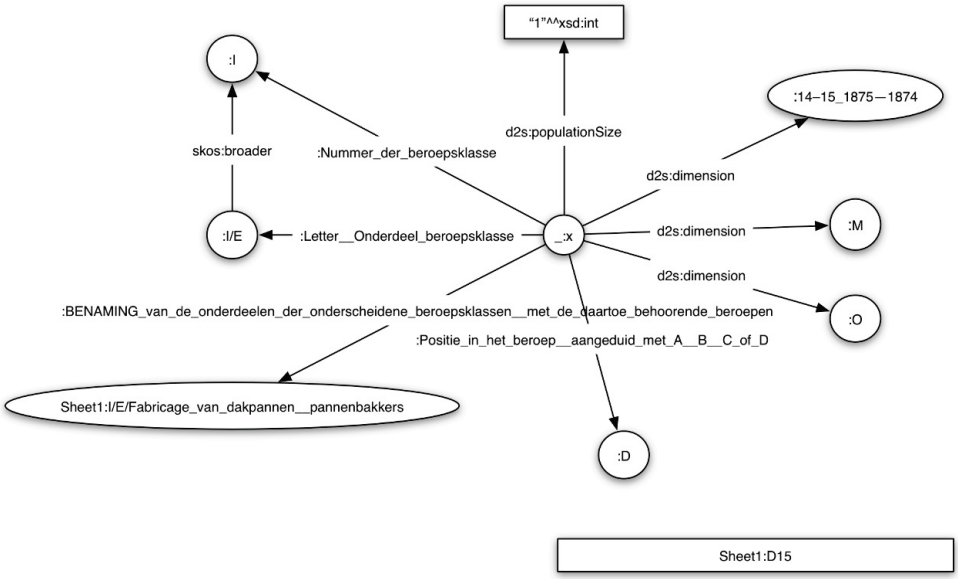


Figure 6.4 - Raw data layer graph

Samples of such separate RDF graphs are shown in Figures 6.4 and 6.6. Figure 6.4 shows what TabLinker produces in the *raw data* layer for *one single data cell*, represented by the central circular node labeled *:x*. This node represents a specific cell of the census tables and its entire environment, namely: the column and row headers that *define* the cell, the data contained in the cell, etc. In this case, the cell contains the value '1' (*"1"^^xsd:int*), and the headers defines the content of the cell as persons of 14 or 15 years old *(:14--15_1875-1874*), being a man (*:M*), being a single (*:O*) and working as roof tile maker (*Sheet1:I/E/Fabricage_van_dakpannen_pannenbakkers*), which is an occupation in the major work category I (*:I*) and subcategory E (*:I/E*) found in cell D15.

The graph representation in Figure 6.4 is only a description of *one* single data cell in the Table. When looking at the representation of Figure 6.4 in a broader context, i.e. the entire census Table represented as a graph, we can easily visualize the structure and complexness of the various census tables which have been converted into RDF. Connecting thousands of cells per Table generates a much larger interconnected graph. For example, in Figure 6.5 we have visualized the census tables of 1869 and 1899 for comparison purpose. As we can note the census of 1889 has a huge increase in the number of data and is much more detailed.

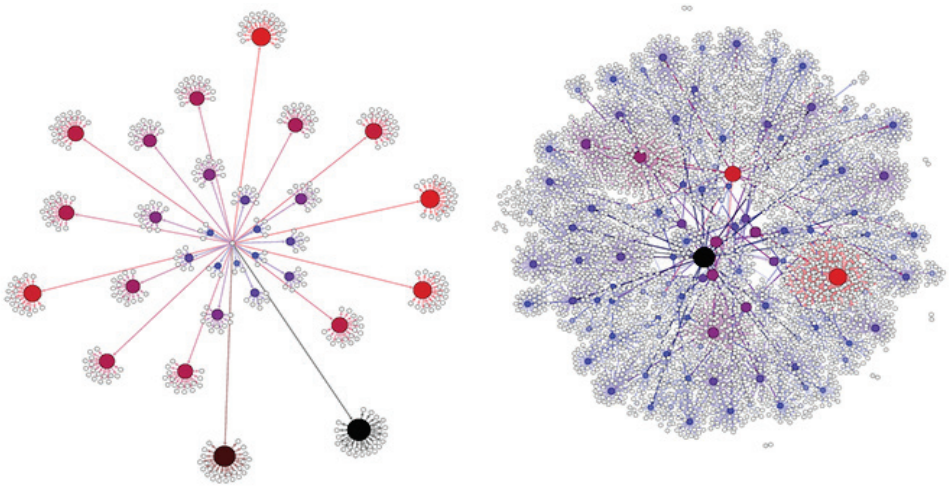


Figure 6.5 - Graph visualization of two different census years (1869 and 1899)

Next to the raw data layer (Figure 6.4), similarly Figure 6.6

specifies an example of an *annotation layer* (labeled *:y*) pointing to a cell with coordinates of the annotation *E663* in the file *VT_1859_01_H1* and the province Table *Noordbrabant*. It includes some metadata, such as the creation date of the annotation (June 21st, 2012), who generated it (somebody represented with the name *TOM*), and a link to the original value (central node in the raw data layer) labeled *:x*. It also contains the flag *1 (see Table 6.1 with flag numbers)*, which indicates that the annotated cell contains an incorrect value (‘*1*’) that should read *10*.

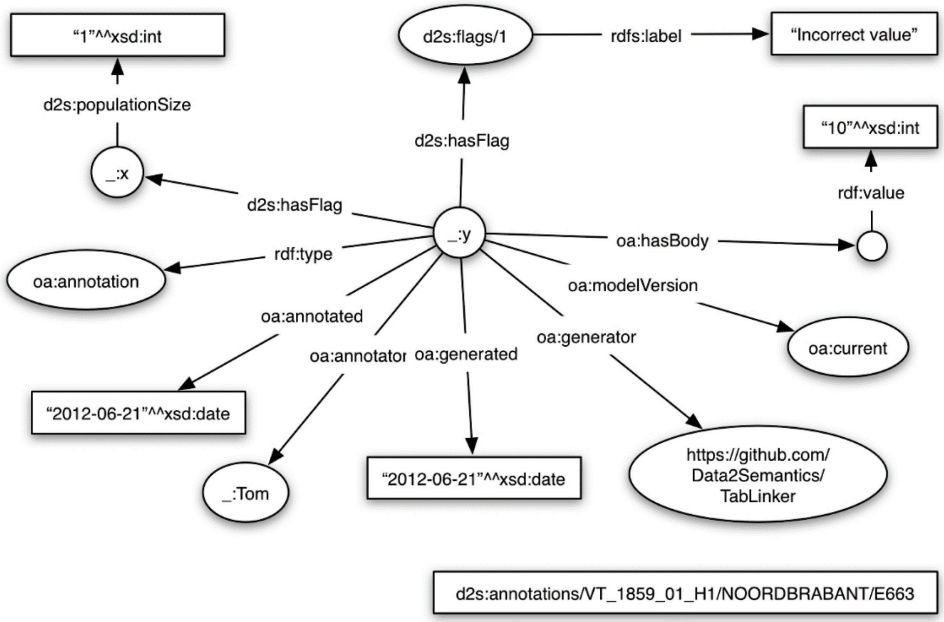


Figure 6.6 - Annotation layer graph

All these graphs are stored in an RDF database called *Triplestore*. From this Triplestore database we will use these raw data layers and build the harmonized database to be distributed to researchers and provide access for live online querying via a SPARQL endpoint[77]. With such an endpoint, users and applications can send census queries in SPARQL to a server holding the dataset, and retrieve results in multiple, known formats such as CSV, HTML, or others.

6.3.4 THE INTEGRATOR – CONNECTING ORIGINAL, RAW AND HARMONIZED DATA

When working with RDF data we need ways to interact with. This practically means that we need to augment it and add harmonizations to it, track the provenance of the changes we make but also provide the trail of the data, process it, produce RDF complaint data, make it queryable, store it in repositories etc. We have developed a data conversion pipeline, where expert users provide the harmonization input and the system ingest this into the RDF Graph with a push on a button. Keeping this process simple and automated allows us to try different approaches very efficiently. We call this system, the Integrator pipeline, see Figure 6.7 on the following page.

[77] http://lod.cedar-project.nl/cedar/sparql

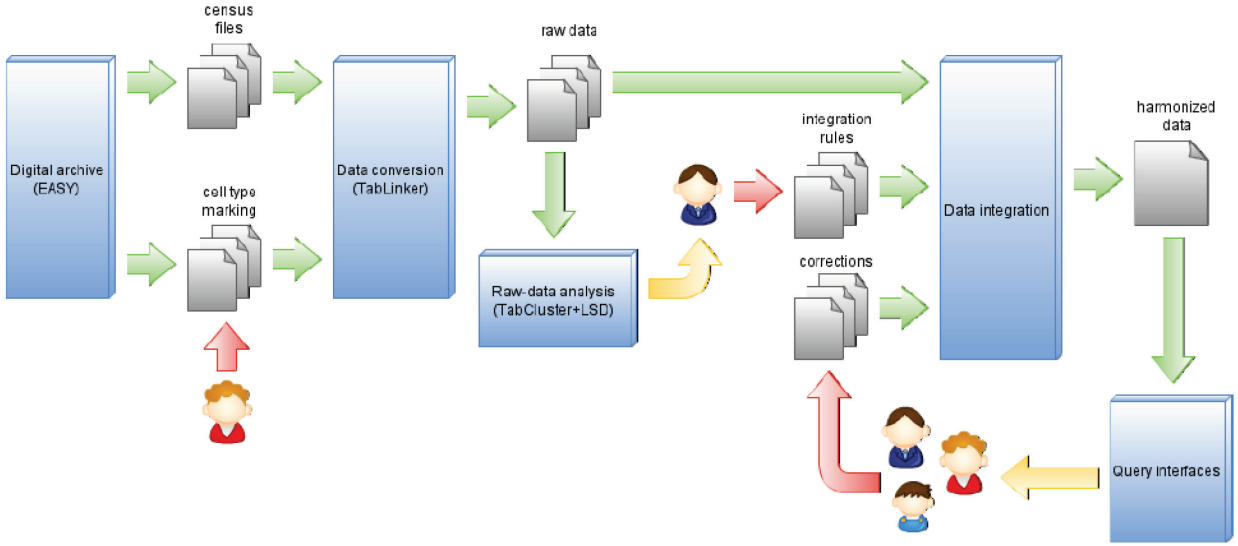


Figure 6.7 - The Integrator - our integration pipeline workflow

Figure 6.7 presents the entire workflow of the Integrator. The Integrator is practically a set of scripts which we automate to connect the *raw* data to *RDF* and the *harmonizations*. Although the harmonization process itself depends heavily on expert knowledge and manual input (see the red arrows), we have automated the majority of the steps (green arrows) which are needed to create new harmonized RDF versions of the data. Next to these there are yellow arrows indicating stages were semi-automated input is given to the expert user.

Looking at the Integrator in Figure 6.7 from left to right, the conversion of the original Excel tables starts with extracting the census tables via the national data archive (DANS) were the tables are stored in an archival system called EASY[78]. Next, after being manually styled (see red arrow) by way of our tool (TabLinker) a new version of the data is created, containing the same content and structures as the original files (i.e the aforementioned 'raw data' layer). In the following stage, the harmonization / integration rules are provided by expert users following a structured approach which we will go into detail in the next chapter. Although the harmonization is based on manual input and knowledge intensive efforts, a set of tools are used to help the expert users in this process. Tools such as TabCluster and LSD dimensions (provides a comprehensive index of statistical variables currently used in other datasets) are used to *assist* the expert user when harmonizing the data (see yellow arrow). The *harmonization rules* in combination with the *raw data* are used to create a harmonized database in the final stage of the Integrator workflow. At the end of the Integrator workflow we have created a feedback

[78] https://easy.dans.knaw.nl/

loop to the expert (end) users in order to process mistakes or corrections when needed (see the yellow arrow at the bottom right Figure 6.7).

Each and every time the Integrator is run, a new dataset is created which we store automatically in a well-known online repository (GitHub[79]), by simply running our integration pipeline. In this repository we store e.g. the original tables, the styled/colored tables of TabLinker, the harmonization input files, RDF graphs (raw and harmonized), our conversion scripts, tools etc.

6.4 PRELIMINARY USES OF THE RAW RDF DATA

After styling and converting all original Excel tables into RDF graphs we can immediately use the raw RDF data in a preliminary stage for various purposes. Although the raw data does not give a solution for the existing problems for longitudinal analysis, the data is now available in its entirety in *one system*. To date, the historical censuses were merely available in the form of unconnected Excel files. All Dutch historical census data (in our raw data layer) are now available via our SPARQL endpoint at one single Web address. Presenting the generated *raw* (RDF) data on the Web via a SPARQL endpoint already enables users and Web applications to retrieve, analyze, and visualize the Dutch historical censuses. Our raw data layer in RDF currently shows the following statistics about our data (produced by TabLinker):

[79] https://github.com/

- 110,585,567 total triples
- 10,272,862 marked cells triples
- 389,132 hierarchical row headers
- 7,960,911 data cells
- 61,110 column headers
- 3,609 row properties
- 2,150 titles
- 1,581,546 row headers, and
- 274,404 metadata cells

At this stage we can already provide some examples of how the layers of our three tier model could interact. For example we have already shown how an annotation influences the result retrieved from the query over a data cell. However the raw data can also be used to analyze the quality of the data, provide insights in the structure and contents of each Table etc. As our database is open at all stages, applications (i.e., applications that use SPARQL endpoints as a data source) can be developed independently by different types of users or other researchers. The SPARQL endpoint (i.e. the RDF database) can be seen, in fact, as an online database 'plug' that any application can use via the Web. This gives researchers, historians, and developers the opportunity to build their own applications (tools) on top of these data. Beyond this, the availability of this dataset as linked data empowers the users to combine its contents with other data hubs on the Web.

With SPARQL, users can merge and remix the data from arbitrary sources on the Web, making the original census dataset richer and capable of answering more with less effort (see Figure 6.1).

Hence, using the raw data layer we are able to retrieving any piece of information of the Dutch censuses. Accessing the raw data mainly allows us to pursue debugging (detecting problems with the data) and harmonization as ongoing work. Practically, querying the raw data enables us to extract the needed variables, assess the quality of the data, identify already common variables across the years for classification purposes, *visualize* them, and detect outliers and inconsistencies (see examples in Figure 6.8, 6.9 and 6.10). For example, by querying for a particular variable (e.g., an occupation, population size, or municipality) across the raw data, we are able to see for which years this variable is present. We visualize this in a simple graph and identify the evolution of the variable across our dataset. Using these practices, we can readily construct the basic branches of our evolution tree (see section 1.3.2 changing variables), that is, we can identify variable creation and extinction which results in an overview of the variables over the years. In our harmonization workflow in the following chapter we show how such 'debugging' practices are needed in one of the first stages of the workflow, i.e. inspection (see section 7.2.2). We use different scripts and visualizations to manage inconsistencies. Several quality checks are provided to the user with regard to the quality of the data. For example, we use outlier detection which displays observations that are numerically distant from the rest of the data. To measure the overall data quality of the raw data we used Benford's Law. Conformance to Benford's Law (Benford 1938) tests whether the frequency distribution of leading digits of all retrieved population counts is the same as the width of gridlines

on a logarithmic scale. Census statistics are well-known distributions expected to obey Benford's Law, and in the Dutch census case the law is met with great accuracy.

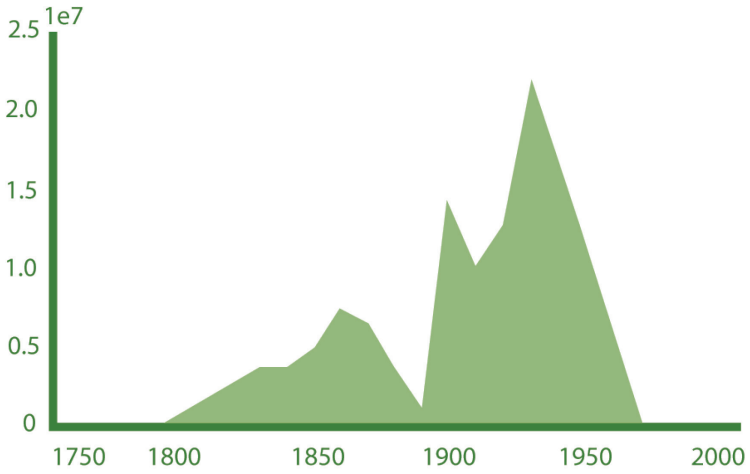


Figure 6.8 - Number of married women over time

Figure 6.8 presents a simple visualization of the total number o- married woman across time. Because we now can analyze the data as a whole (in one system) these type of visualizations are used to detect problems in the raw data. As we can see for the year 1899, there is a clear decline regarding the number of women. Obviously this means that we have to go back to original sources and detect why there is missing data for this specific year.

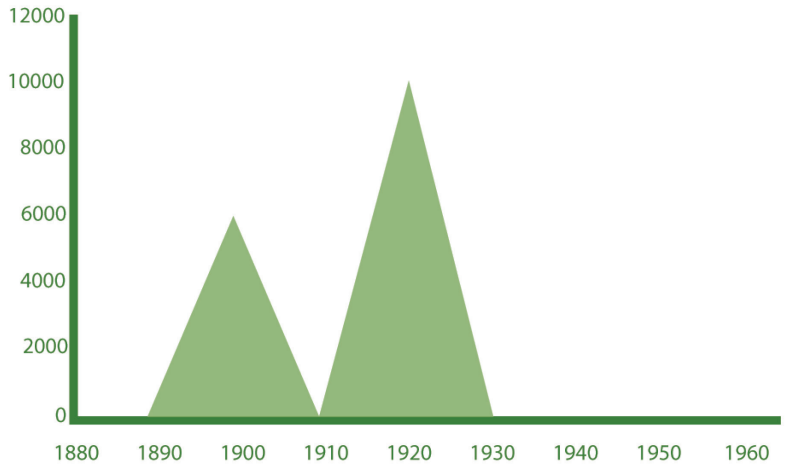


Figure 6.9 - Number of teachers (HISCO 13490) over time

Using the HISCO classification system we have mapped the occupations contained in the original census tables, in a preliminary and straightforward method to HISCO codes. In this example, looking at the *number of teachers* over time (Figure 6.9) we clearly see missing data for the census of 1889, 1909, 1930 and beyond. We use simple visualizations like this to detect errors and improve upon our data by going back to the original sources and verifying the missing results.

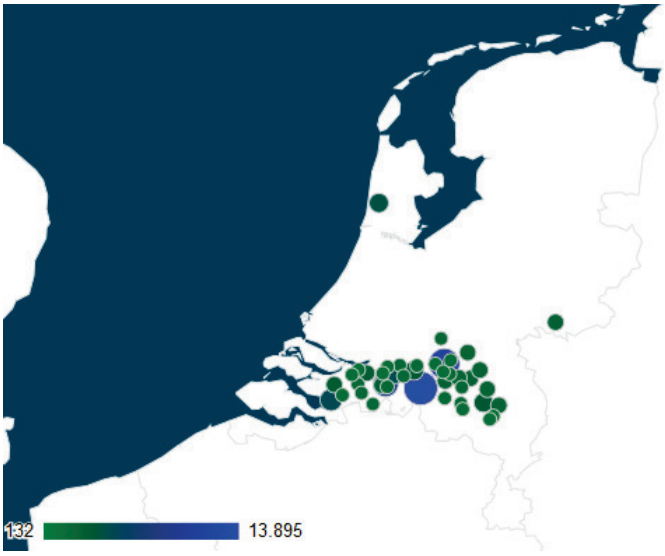


Figure 6.10 Displaying the municipalities in the Province of North Brabant on a map for outlier / data error detection purposes

In Figure 6.10 we have plotted the number of males and females for the province of North Brabant. Here we use visualizations on a map to detect mistakes and outliers in the data. In this example we clearly find several municipalities which fall outside this province and need to be mapped to the correct municipal codes.

6.5 CONCLUSION

In this chapter we have presented a model for harmonizing (historical) aggregate data where we have separated the original data from the annotations and harmonizations rules. This 'three tier' model allows us to gradually harmonize the census data whilst not changing the original sources. We next present a generic and straightforward way for converting the unconnected and dissimilar census tables to RDF format. Currently, we have progressed to the point of converting all raw datasets from the census Excel spreadsheets into an RDF triple dataset. In the next chapter we build on this raw data layer in order to build our harmonizations.

By separating the actual harmonization from the original source data we make it possible to follow a source-oriented harmonization approach. Such an approach allows different views on the data, the option trace back the data to the original sources (on a cell level), and provide accountability as to who made changes. These advantages we provide are key requirements in historical research. We next showed the preliminary uses of having everything into one graph system. Besides various visualizations used to identify outliers or mistakes in our data we also have measured the overall quality of the data by applying Benford's Law on our dataset, to which our data conforms.

7. SOURCE-ORIENTED HARMONIZATION OF HISTORICAL CENSUS DATA: A FLEXIBLE AND ACCOUNTABLE APPROACH IN RDF

While many (successful) efforts have been undertaken by researchers to harmonize historical census data, a lack of a generic workflow thwarts other researchers in their endeavors to do the same. In order to compare historical census data over time, a common process currently often loosely referred to as; *harmonization,* is inevitable. The process of harmonization becomes even more challenging when dealing with aggregate data. Current approaches whether focusing on micro or aggregate data mainly provide specific, goal-oriented, solutions to solve this problem. The nature of our data (i.e. historical), calls for an approach which allows different interpretations on the data and preserves the link to the underlying sources at all times. To realize this we need a flexible, bottom up system which allows us to iteratively discover the peculiarities of our data and provide different interpretations in an accountable way. In this chapter we propose an approach which we may refer to as, source-oriented harmonization. We use the Resource Description Framework (RDF) from the Semantic Web as the technological backbone of our efforts. In this chapter we propose a structured source-oriented harmonization workflow which can be used to iteratively explore and harmonize data such as the historical censuses in an accountable way. By doing so we aim to make the task of harmonization less of a random process which often focusses too heavily on standardization alone. We aim to make this, until now vague, process more graspable for others and stimulate similar efforts beyond census data. In this chapter we continue and look

deeper into harmonization layer of the model we proposed in Figure 6.2.

"Data do not give up their secrets easily. They must be tortured to confess"

Jeff Hopper, Bell Labs

This chapter consist of original work and two published articles (1) Ashkpour, A., Mandemakers, K., Boonstra, O. Source Oriented Harmonization of Aggregate Historical Census Data: a flexible and accountable approach in RDF. Historical Social Research / Historische Sozialforschung, 41(4), pp. 291-321. (2016). In this work I was the main author and the developer of the source-oriented harmonization workflow we present in this chapter. Next to the methodology and harmonization approach, the data produced, is a result of extensive iterations and knowledge intensive harmonization input. (2) Meroño-Peñuela, A., Ashkpour, A., Guéret, C." Proceedings of the 2nd International Workshop on Semantic Statistics (SemStats 2014), ISWC 2014, Riva del Garda, Italy (2014). A modified version of this article is used as a sub section in chapter (7.2.4). My contributions were related to the development of tools for bottom up classification building and the actual creation and curation of the expert classification system used in this article. This classification system was used as the 'golden standard / ground truth' against which we compare the results of the automated classification suggestions. (3) Original text in 7.4 and 7.5.

7.1 INTRODUCTION TO THE PROBLEM

Throughout history, the main goal of censuses has been to collect information about a nation's population characteristics. As censuses are meant to accommodate the *information needs* of governments and societies, changing circumstances will require different questions and data. Questions and purposes therefore change for each census. This principle is very well reflected and inherent in understanding the changing nature of historical censuses. These changes are valuable snapshots of our history (Higgs 1996) and are embedded in the very structure of the census itself, resulting in changing questions, variables, classifications, structures and processing methods over time.

From census to census we find changing questions, enumeration methods, variables, values, classifications etc. all hampering longitudinal analysis of the data. The diversity in data formats, structures and content of historical censuses calls for a unified system. *Harmonization* is therefore a prerequisite in order to do *any* type of longitudinal research. However the harmonization process differs for micro and aggregated data. The main difference is that aggregated data introduces more ambiguity. Whereas with micro data one is able to build classifications systems, variables etc. according to one's need, aggregated census data needs to introduce estimation schemes to achieve results which can be used for comparable research.

As a consequence, when dealing with aggregate data we do not know beforehand which harmonizations are the most optimal choices. It is a process of trial and error, also depending on the research in question in mind.

For this reason, it is necessary to have flexible systems which enable us to create different harmonizations on the same variables in an iterative way. Current census harmonization practices and models are not designed in such a way and usually do not (easily) allow different views on the data. These, mostly micro data practices, result in only one version of a newly categorized and classified dataset. Because of this, current efforts therefore lean more towards 'goal-oriented' methods, where the users of the data depend and are bounded to choices and interpretations which have been set before (Cameron and Richardson 2005; Thaller (1993). According to Greenstein's (1989) definition, the source-oriented approach should allow two main requirements. Namely that, the same source is handled differently in various stages of historical research and that the uses of sources vary over time.

The source-oriented approach is the preferred method in historical research. Being able to refer to the original sources and allowing different interpretations on the same data is an important requirement in this field of research. However, thus far the harmonization of historical census data is based on goal-oriented methods. What we need are more source-oriented data processing methods, which do not force the historian to make a decision on which methods to be applied at the time of the database creation (Boonstra, Breure and Doorn 2004, Thaller 1993). With aggregate data we need a flexible bottom up approach which allows a learning experience, to iteratively test and provide different harmonizations in order to deal with the ambiguous nature of such data. It is an approach that we refer to as "source-oriented harmonization".

In this chapter we explain how we implemented a source-oriented, structured harmonization approach with the Dutch aggregated historical census data, using Semantic Web technologies. We propose an iterative harmonization workflow in RDF (Resource Description Framework) which makes different interpretations possible without losing track of the original data. We provide tangible links from the harmonized data to the original sources upon which the harmonizations are based. In doing so accountability is guaranteed at all stages. In the following sections we start by looking at census harmonization in general and describe its challenges. Next, we go into the details of our workflow and explain each step of our suggested approach and how we used this to gradually build harmonized tables in the context of a pilot project. We end with a discussion of the results and the wider impact of our source-oriented harmonization system.

7.2 THE HARMONIZATION WORKFLOW

In order to explore the possibilities of publishing the original and harmonized data of the Dutch historical censuses in the Semantic Web, using RDF, we developed a pilot to test our methods and workflow. For this pilot we focus on a subset of the censuses that contains the number of inhabitants and dwellings for each locality and municipality. We selected these so-called *Local Division* tables for the census years 1859, 1869, 1879, 1889, 1899, 1909 and 1920. The local division tables provide insights on both generic statistical data such as population totals across the different areas but also in depth distinctions on the different statuses assigned to

the population. Providing a harmonized version of this data, the state of the nation can be studied on abstract levels such as the total number of inhabited houses, houses under construction, houseboats or the number of males/females. It will also be possible to ask detailed questions such as 'the total number of people counted in monasteries in the centers of small towns for each province'.

To harmonize these census years we have chosen to start with comparable subsets of the data. We initially started with 1869-1879, 1889-1899 as these share similar classifications. By defining these, and gradually adding additional years and classifications we included three more census years to the data. The data of these seven census years are currently stored in 60 heterogeneous Excel tables. For some years there are large Excel files containing different tables (sheets) per province, for other years we have smaller Excel files (tables) for each province separately. The number of tables and measure of detail differs widely between the different censuses while containing different types of variables on different geographical levels.

Figure 7.1 presents a scheme of the source-oriented harmonization workflow we have developed. The points of departure for the construction of this scheme were the following principles: 1) The workflow should be applicable to other similar datasets, 2) The workflow should allow systematic testing and feedback loops in all stages, 3) The raw data should never be changed, 4) Since data as complex as historical censuses cannot be harmonized in one try, the workflow must allow an iterative processes of trial and error, 5) Different interpretations on the data should be allowed and 6) we should always be able to point to the

underlying (original) source data (i.e. Excel files, images, books etc.).

The first step in the workflow (Figure 7.1) consists of the (source-oriented) **conversion** of the original Excel tables into RDF format. Once the data is converted, we create a *1:1 Graph database* (an important point of takeoff in our approach) which we use to build our harmonizations on. The second step is the **inspection** stage of the data. During this stage we try to get a better understanding of our variables and values by directly querying the newly created RDF database. The first feedback loop of our workflow starts here**.** The third step of the workflow is the **standardization stage** where we actually make harmonization decisions on how to define the different variables and values uniformly over the years. After standardization we move on to **classification** stage and put the numerous various variables and values into meaningful groups. During this stage we create internal bottom up classifications and make use of external classification systems wherever possible, whilst enriching the web with our census specific systems (see feedback loops to 'external classification systems and variables'). The next part of the workflow is the **variable/value creation section**, where we actually create (missing) variables and values to fill in the gaps of our tables and bridge between the different censuses. Depending on the needs after standardization, this stage could be applied prior, after or simultaneous with the classification stage. For example, it may be that some variables need to be grouped into other variables to make meaningful classifications. The finishing touch of the workflow is the **testing** of all procedures. We 'test' the produced data extensively after each stage of our workflow in an iterative manner by querying the database and creating

intermediate tables until a certain degree of quality is reached. Now harmonized and tested, in the **create dataset** stage we produce different types of tables for researchers. In the following sections we go into the details of each stage and how we worked towards producing a harmonized dataset for the Dutch *Local Division* tables of the census.

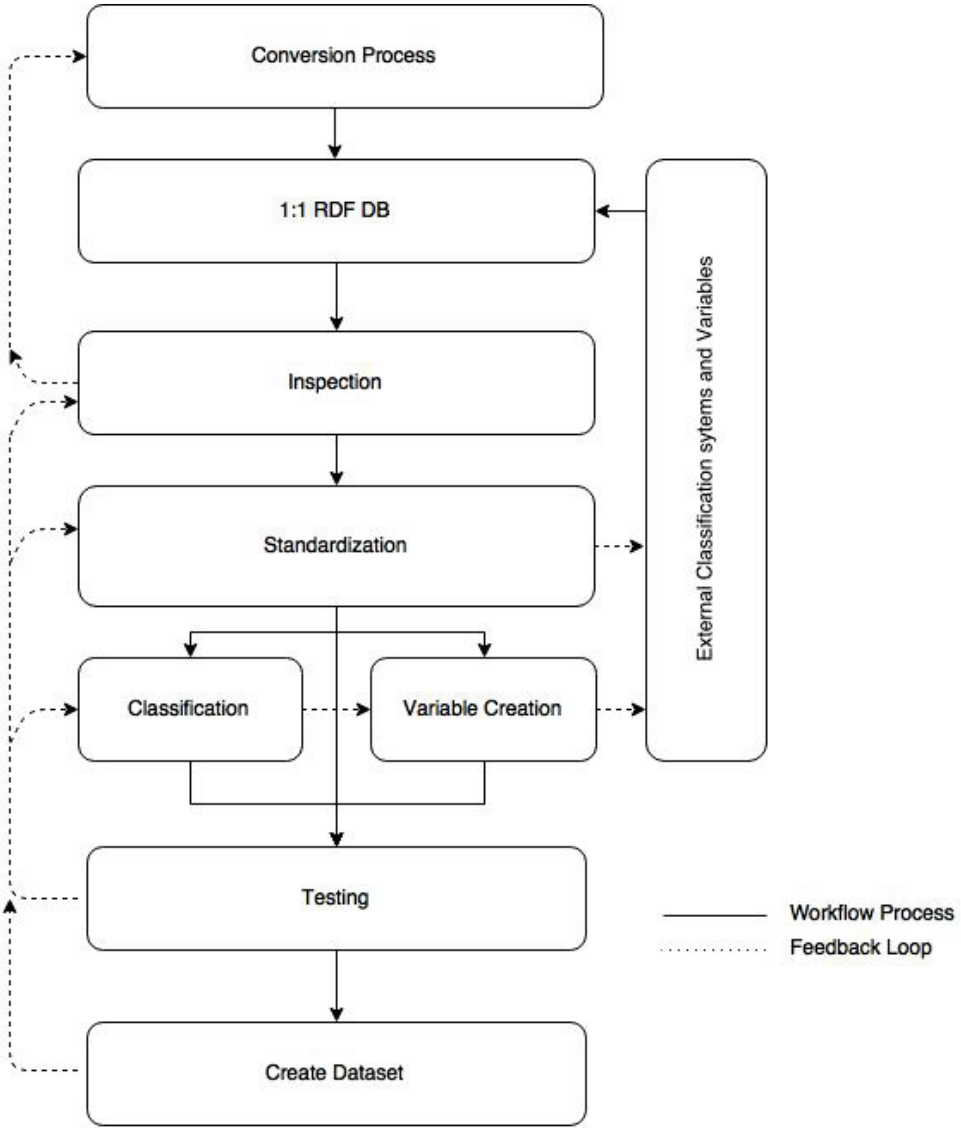


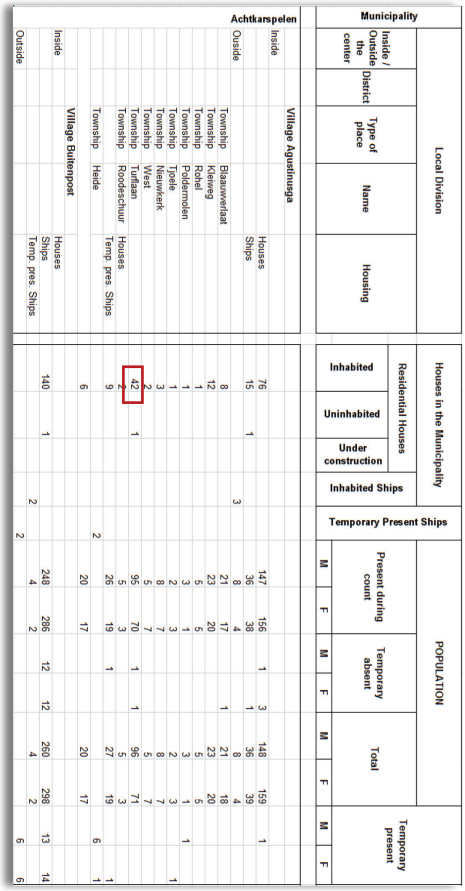
Figure 7.1 - Source-oriented Harmonization Workflow of aggregate Historical data

7.2.1 CENSUS DATA IN RDF: CONVERSION AND 1 ON 1 MODEL

The first step in our harmonization workflow is to convert the data and its original hierarchies and structures from the Excel sheets, in which they were stored in 1997, into a RDF database. Providing the challenges faced when trying to harmonize historical censuses, we recognize the need for a flexible and downright bottom up approach. The premise of our source-oriented approach builds on the notion that the underlying dataset should be converted into an RDF database *without* making *any* decisions on how to model the data beforehand. This means that we represent the historical data sources as one to one copies in RDF. By converting the data to RDF we gain the advantage that we are now able to query the census tables as a whole. This allows us to explore, discover, try, fail and try again in order to learn the data with all its peculiarities before committing to a certain interpretation. By not dictating a predefined model in our harmonization workflow, any similar dataset could follow our approach and apply the workflow to their own data.

Currently, the application of Semantic Web technologies is being advocated in different historical fields (Meroño-Peñuela et al. 2015a). Different types of historical data are being converted to RDF using a variety of tools and methods. In order to move towards a database in RDF, an appropriate *RDF model* should be used. Depending on the *type* of data (e.g. textual, statistical, structured etc.) different models are available. In the case of *census* data we used *RDF DataCube*, the *standard* in the Semantic Web for modelling 'multi-dimensional statistical data' (Cyganiak et.al 2014). This model is based on the *SDMX* cube model and ISO

standards for the exchange of statistical data. In order to convert our Excel tables (see Figure 7.2) we used a very straightforward tool (TabLinker[80]) to convert the different census tables into RDF DataCube compliant data.



Municipality						Houses in the Municipality				POPULATION							
Local Division						Residential Houses			Inhabited Ships	Temporary Present Ships	Present during count		Temporary absent		Total		Temporary present	
Housing	Name	Type of place	District	Inside / Outside the center		Inhabited	Uninhabited	Under construction			M	F	M	F	M	F	M	F
Achtkarspelen																		
Inside																		
Village Agustinusga																		
Houses						76					147	156	1	3	148	159	1	
Ships						15	1		3		36	38	1	1	36	39		
Outside											21	4			21	4		
Township	Blauwestaat					8					21	17	1		21	18		
Township	Kleiweg					12					23	20			23	20		
Township	Rohel					1					5	5			5	5		
Township	Poldermolen										3	1			3	1	1	
Township	Tpele					1					2	3			2	3		1
Township	Nieuwkerk					3					8	7			8	7		
Township	West					2					5	7			5	7		
Township	Turfdaan					42	1				95	70	1	1	96	71		
Township	Roodeschuur					2					5	3			5	3		
Houses						9					26	19	1		27	19	1	
Temp. pres. Ships										2								1
Township	Heide					6					20	17			20	17	6	1
Village Buitenpost																		
Houses						140	1		2		248	286	12	12	260	298	13	14
Ships											4	2			4	2		
Temp. pres. Ships										2								6
Outside																		

Figure 7.2 - Original Excel Table with the number of inhabitants and houses per geographical unity for the census year of 1889 (column and row headers are translated from Dutch)

[80] https://github.com/Data2Semantics/TabLinker

Figure 7.2 is an example of an (transcribed) Excel Table to illustrate the *structure* and *contents* of our (source) data. This Table contains the number of inhabitants and houses per geographical unity for the census year of 1889. The Figure shows how the different numbers in the tables are connected to *multiple* row and column headers. These headers contain different types of hierarchical variables and values. For example the highlighted cell with the number "42" is connected to the 'township' called 'Turflaan' belonging to a 'village' called 'Augustinusga' in the municipality of 'Achtkarspelen', in a place 'outside the center', presenting the number of 'inhabited residential houses'.

In our conversion to RDF we preserve the original structure and dependencies between the different variables. Using the TabLinker tool, we took advantage of the structured layout of the Excel tables and define the different areas where the numbers and variables/values are contained. Figure 7.3 shows an example of the same Table as in Figure 7.2, but now styled using the definitions provided by Tablinker. We will use this table as an example throughout the different stages of the workflow in the following sections of this chapter.

Municipality					Houses in the Municipality						POPULATION									
Local Division					Residential Houses			Inhabited Ships	Temporary Present Ships		Present during count		Temporary absent		Total			Temporary present		
Housing	Name	Type of place	District	Inside / Outside the center	Inhabited	Uninhabited	Under construction				M	F	M	F	M	F		M	F	
Achtkarspelen																				
Inside		Village																		
	Houses	Agustinusga			76						147	156	1	3	148	159		1		
	Ships				15	1		3			36	38		1	36	39				
Ouside											8	4			8	4				
		Township	Blauwverlaat		8						21	17		1	21	18				
		Township	Kleiweg		12						23	20			23	20				
		Township	Rohel		1						5	5			5	5				
		Township	Poldermolen		1						3	1			3	1		1		
		Township	Tjoele		1						2	3			2	3				1
		Township	Nieuwkerk		3						8	7			8	7				
		Township	West		2						5	7	1		5	7				
		Township	Turflaan		42	1					95	70	1	1	96	71				
		Township	Roodeschuur		2						5	3			5	3				
	Houses				9						26	19	1		27	19				1
	Temp. pres. Ships	Township	Heide						2											1
					6						20	17			20	17		6		1
		Village	Buitenpost																	
Inside	Houses				140	1		2			248	286	12	12	260	298		13		14
	Ships										4	2			4	2				
	Temp. pres. Ships								2											6
Outside																				

Figure 7.3 - The same Table as in Figure 7.2 but now styled with our conversion tool

After styling and converting the tables we create a RDF database using the Tablinker scripts. This 1:1 RDF database contains the same content and structures as the original Excel files. We call this the 'raw data' layer.

Figure 7.4 is a graphical representation of the Excel tables after they have been transformed to RDF. Notably this example is just *one* RDF graph representing *one* Excel Table, after the conversion to RDF we create a graph for *each* of our tables. Creating links between these graphs is what allows us to query the data across time. The raw data layer together with several queries and scripts are used to *assist* the harmonization process in the next stages of our workflow. By harmonizing the data we aim to build links between the different graphs and query them in longitudinal ways.

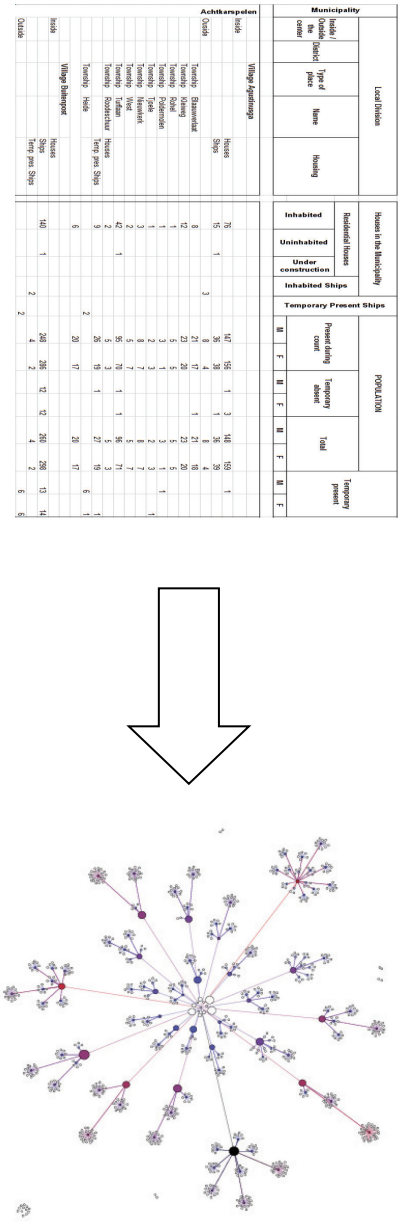


Figure 7.4 - Graphical Representation example of the Excel tables in RDF

It is very important to keep in mind that during this stage of the workflow, under no circumstances the original data should be touched, even when obvious mistakes are spotted. By doing so we are always able to reproduce the provenance (W3C, 2015) of our actions, back to the original source material. Data or transcription errors and ambiguities will be dealt with later in the process, by structurally going through the different steps of our workflow. Errors made in the conversion are dealt with by improving and running our system again. We provide in depth technical descriptions for those interested in understanding and setting up a similar workflow (Meroño-Peñuela et al. 2015b).

7.2.2 INSPECTION

The next step is the inspection of the data in a semi-automatic way. Out of the enormous pile of raw data in RDF format, we now need to identify the different classifications, variables, values etc. contained in the original censuses. In other words, before we can define the variables we first need to know what we have. At this preliminary stage we can already analyze the raw data as a whole to provide *insights* for the harmonization itself. So, while staying true to the source-oriented approach (Boonstra, Breure and Doorn 2004; Thaller 1993; Cameron and Richardson 2005), we have created a database which we now can use to query in order to get statistics about the landscape of the historical censuses. We can ask questions such as e.g. what are the different variables and their values, which ones are the most frequent used (baseline statistics), how are these variables related, can we find similar classification systems, do we need all literals to define a variable ? etc.

During this stage we clearly need to define the scope of the data which we want to harmonize. Although changing definitions is a known hindrance to historical census harmonization, there *are* certain periods in which the census share *common* characteristics such as the same classifications, variable, values, questions, structures etc. By starting with harmonization of censuses which share similar characteristics, we create general rules and practices which can be extended to the entire dataset. For instance, in the case of the *Local division* tables we can identify that there are three subgroups of censuses which use similar classifications i.e. 1859–1879, 1889–1899 and 1909-1930. This is very useful as the harmonization input itself is heavily dependent on expert knowledge and human input. Therefore, not exposing the data to the experts as one big dump, makes it easier to get a better grasp on the data when analyzing it as a whole (Slavakis, Giannakis and Mateos 2014).

After similar subgroups have been identified and the scope is set (by using expert input), it is time to start looking at its content. The first major step in the inspection process is to make frequency distributions of the different variables and values to see what we actually have across the years. We do this by directly querying the raw data layer. To get a clear idea we have to look at this in twofold. First, we make *univariate frequency lists* of the raw variables and values in order to create data driven vocabularies. Second, we create *hierarchical frequency lists* to understand the mutual connections between variables and how these are hierarchically situated in the tables. As we will illustrate further on, the context and *relationship* of the variables are key to the

understanding and creation of formal descriptions of the data. For example, where a frequency list (Table 7.1) would merely give us an overview of the variables and values which occur most often, a multivariate hierarchical frequency Table (Table 7.2) shows how the terms are connected in the original tables. This helps us to understand its context, and by this, the nature of the variables and its values.

Literal	#
Males	8981
Females	8721
M.	654
F.	607
Temp. Present	4506
Temporary Present	2151
Pop	1015
Population.	9647
Population	2458
Legal Present	2412
Leg. Present	894
Legally Present	2452
Factual Present	5853
Total.	2545
HouseBoats	5482

Table 7.1 - Sample of a Frequency List of 'raw terms' in the original tables and directly generated by querying the RDF Graph

Year	Variable 1	Variable 2	Variable 3
1869	Temporary Present	F	
1869	Temporary Present	M	
1889	Temp. Present	Males	
1889	Population	Males	Legally Present
1879	Population	Males	Total.
1899	HouseBoats	Temporary Present	
1879	Population	Females	Total.
1899	Population	F.	Legally Present
1899	Population	M.	Legally Present

Table 7.2 - Flattened list example of the hierarchies among the variables in a census Table, directly generated from the RDF Graph

The examples in Table 7.2 and Table 7.2 show the results of the data inspection stage (note these are samples for illustration purposes only). Table 7.1 is a simple frequency list of the literals (strings) used most often in the census. Here the variables are viewed as independent, which they are clearly not. Table 7.2 presents the same terms as in Table 7.1 but now in relation with each other. In this *hierarchical* view we have flattened for example the variable combination 'Temporary Present' and 'Males/Females' (to represent the original hierarchy, see Figure 7.2). Simply looking at the frequency list (Table 7.1) makes it difficult to make sense of the meaning and context of the variables; by considering the original hierarchies of the variables, we now see for example that term "Temporary Present" is connected to both

"Sex" (a demographic variable) and "HouseBoats" (a housing type variable). By providing this information to the expert user we assist them in the process of creating distinct formal definitions. For the queries we have used to extract these literals from the RDF tables, see our website www.censusdata.nl. Our example clearly show the importance of maintaining the connections between the variables as contained in the original tables. The *combination* of these variables is what allows us to create valid queries on the data (will be elaborated the following section under 'variable mappings').

As a result, during the inspection stage we focus in the first instance on identifying subgroups of censuses which share similar characteristics. Within these subgroups we focus on the most frequent and/or important variables in the censuses and their relations. By doing basic analysis we are able to identify a set of literals which account for the majority of the variables in the census tables. After this is done, we focus on the details and specificities of less frequent variables. Using this information we are able to create, in a semi-manual way, a variable overview across the years which will serve as the *input* for the next stage of the harmonization process i.e. standardization. This overview therefore also provides insights in the order in which the variables will be processed, i.e. using the frequency lists and expert input we have categorized the variables according to their level of difficulty. During the subsequent stages of our source-oriented workflow we systematically come back to the inspection stage to identify new problems and to improve the standardizations, classifications and variable we have created along the way.

7.2.3 STANDARDIZATION

In our source-oriented harmonization approach we first have converted everything 'as is' into one RDF system. This means that the variables are still only accessible by their own labels (literals). To allow longitudinal analysis we still have to standardize each and every single variable and value in this new RDF database. *Standardization,* is the first harmonization stage in our workflow where we have to decide on how to make the data uniformly accessible over the years. During this process expert knowledge about the source data is key in assigning meaningful definitions and mappings. In this section we describe the four different elements of our standardization process. We start with a *selection* of variables and values to standardize, next we *formally define* the identified variables and values. Once defined we describe the *grouping* of the values and we finish with illustrating the importance of maintaining valid variable *mappings*. This structured standardization procedure enables us to access all the different variables and values uniformly over the tables and extract all relevant data.

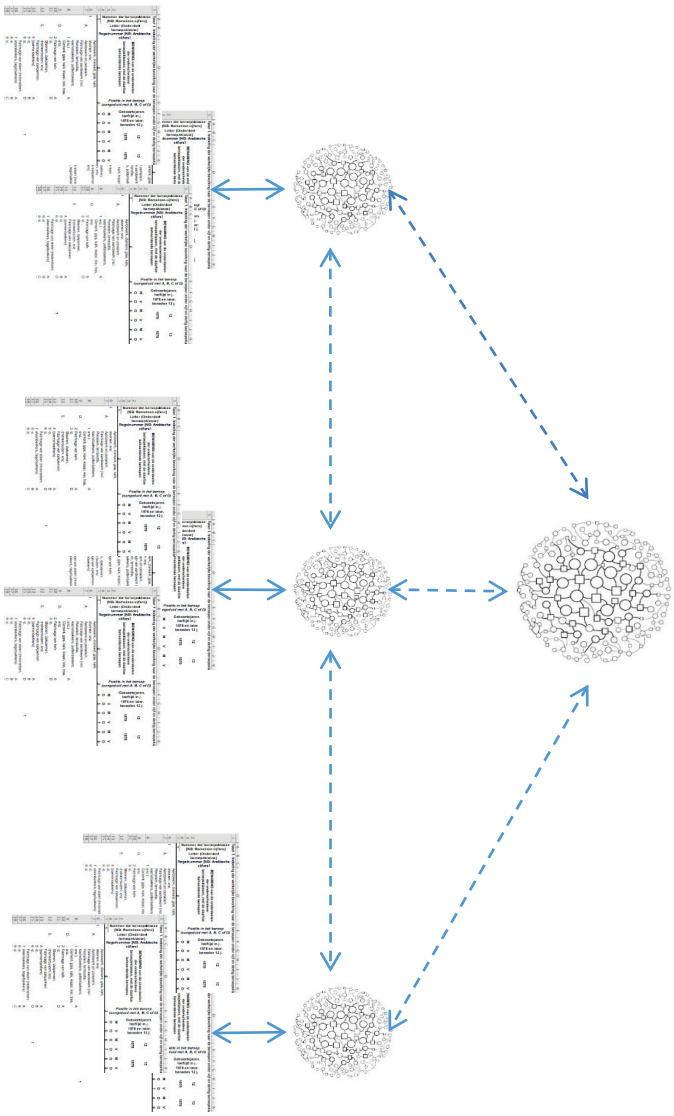


Figure 7.5 - Illustrating the need for harmonization and connecting the different RDF graphs

The dotted lines in Figure 7.5 illustrate the need for harmonizing the raw RDF graphs in able to interconnect them. By going through the different steps of the workflow we fill in dotted lines and provide the links between these graphs and produce a harmonized census database. The first step in this process of interconnecting the raw graphs is *standardization*.

UNDERSTANDING THE DATA STRUCTURE: A FIRST SELECTION OF THE VARIABLES

Figure 7.6 presents a census Table which describes the number of inhabitants and houses for a given year. Each data cell (number) in this Excel Table is connected to various column and row headers of the census Table. These headers represent the multiple dimensions of our RDF model. During this stage we have to determine the meaning of all literal values such as "Achtkarspelen", "Uninhabited", "Males", "Huizen", "Temporary Present" etc. and all their variations. We use the input from the *inspection stage* to build 'bottom up' standardizations. Based on this we *first* select those variables that are necessary to define a number in the table. This is what we call the '*minimum required definitions'*.

"Not everything that can be counted counts, and not everything that counts can be counted"

William Bruce Cameron

In our RDF model the data cells / numbers are the central node. This principal means that the RDF graphs are built by connecting all the dimensions (variables and values) contained in the headers to their corresponding numbers. In Figure 7.6 we see an example of this and show the eight different row and column headers that are connected to the (bold) number we are interested in, i.e **113**. These headers or dimensions are indicated by arrows in the Table. The number 113 refers to the total number of temporary present males in the municipality of Achtkarspelen, thus for this number only three out of the eight dimensions are needed to define it, i.e. *Municipality*, *Temporary Present* and *M*.

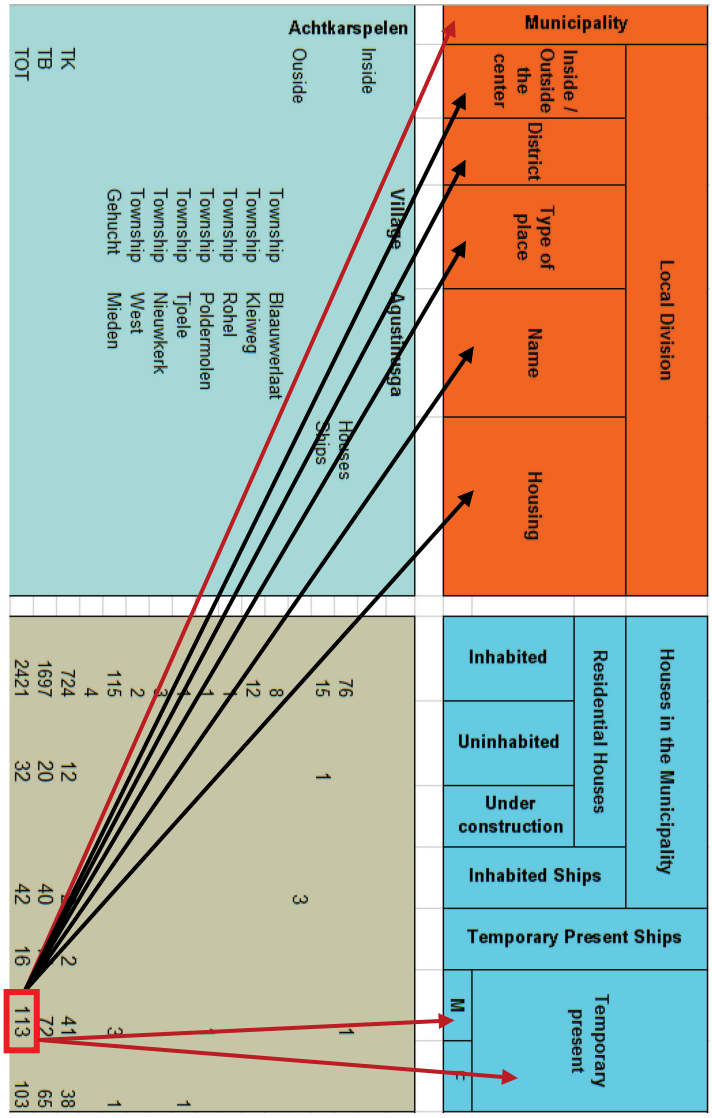


Figure 7.6 - Excel Table highlighting the different dimensions which are related to the bold number

We therefore (first) provide standardizations for:

- Municipality – "Achtkarspelen"
- Residence Status – "Temporary Present"
- Males – "M"

By standardizing these three variables, *in combination*, we are able to retrieve this *specific* (113) number from our tables. We transfer the *totals* (TK, TB, TOT) to RDF and standardize them for comparison purposes but purposefully ignore these values in the query process in order to avoid over-counting as we create our own totals. This is needed because the original totals are not always correct or sometimes even missing. Moreover by creating totals using all the lower sub-values we can break down and study a total in case of wrong values and i.e. identify the specific cell which is wrongly standardized. The more we define and standardize the more specific we can target a data cell in the tables. For example, to get the total number of '*males'* which are '*temporary present'* in a specific '*district', 'outside the center' or in a certain 'housetype'* of that 'municipality', the lower geographical areas need to be defined in addition to the municipality. The iterative nature of our workflow allows us to start the standardization at more abstract levels and focus on the specificities and details in later stages. This is necessary to rise above the data deluge problem so the experts can get a better grasp on the data. To keep track of our progress we frequently produce statistics to see how much of any given Table is defined and what is still left.

PROVIDING FORMAL DEFINITIONS

Building on the input from the inspection stage and by identifying the *minimum required definitions* we provide standardized terms for the given literals in a structured way. During this process we enrich the literals with standardized terms. By doing so we are able to access the data across time and space using a *common* vocabulary. This means that we consistently assign standard definitions or codes to all possible variations of a given variable or value. See Table 7.3 for an example of how we use the input from the inspection stage to standardize the terms in a structured way.

1869				
Original String	**Standardized**	**Original**	**Standardized**	**Formal Expert Definition**
Total	Legally Present	M	Males	Legally Present Males
Total	Legally Present	F	Females	Legally Present Females
Present during count	Actually Present	M.	Males	Actually Present Males
Present during count	Actually Present	F	Females	Actually Present Females

Table 7.3 - Using the frequency list and flattened hierarchical view formal definitions are given by expert users of the data.

Each line in this Table has to be seen as a possible variable *combination* (based on the original hierarchies, whereby the original terms are translated into English in Table 7.3). In order to query for all the dimensions and their combinations, the different values first need to be defined separately. The blue columns represent the original terms/literals in the tables (extracted during the inspection stage). The last column is the *formal definition* given by the expert user and the yellow columns are the standardized terms given by us, based on the formal definition. We follow this approach to structurally standardize all the literals in our raw RFD dataset. By formally defining the dimensions related to the numbers in the excel files we are able to retrieve any *specific numbers* from the census tables.

PUTTING VALUES INTO STANDARDIZED VARIABLES: GROUPING

At this stage of the standardization process the literals are formally defined and standardized, but they still are not grouped into meaningful variables or domains. For example, the literals Male and Female are now standardized and accessible uniformly across the tables but what are males and females? What do 'Temporary Present', 'Actually Present', 'Legally Present', 'Houseboats' etc. mean? In order to give them meaning we need to put them into standardized variables, i.e. variables which have been created by ourselves. In our example, we assigned our values to three standardized variables. For example, we attached the Male and Female values to the standardized variable *Sex*. The four different statuses given to persons or housing types are defined as *ResidenceStatus*. Finally we have created a standardized variable for

the different *HouseTypes* (houses, wagons, houseboats, carehomes etc.). These standardized variables *together* are what allow us to reconstruct the original variables and values during the querying process when combined and reshuffled, as will be explained below. For example we could now make a query which gives us: Municipality: Amsterdam, ResidenceStatus: TemporaryPresent *and* Sex: Males to recreate the variable Temporary Present Males in the Municipality of Amsterdam.

MAPPINGS

In order to (correctly) use the standardized variables and query the data, one last important step remains. An expert user may know which combinations of variables are possible, but others may not. Merely providing a query endpoint where users can enter queries does not work if they don't know what to query for *and* more imporantantly, which combinations are possible. In this final stage of the standardization, we need to provide and maintain valid mappings to the variables in order to guide users in making (correct) queries. The result of the different standardization stages are represented in our standardization model below as showed as connections in Figure 7.7. This Figure shows the standardized variables, the standardized values *and* how they are related to one another.

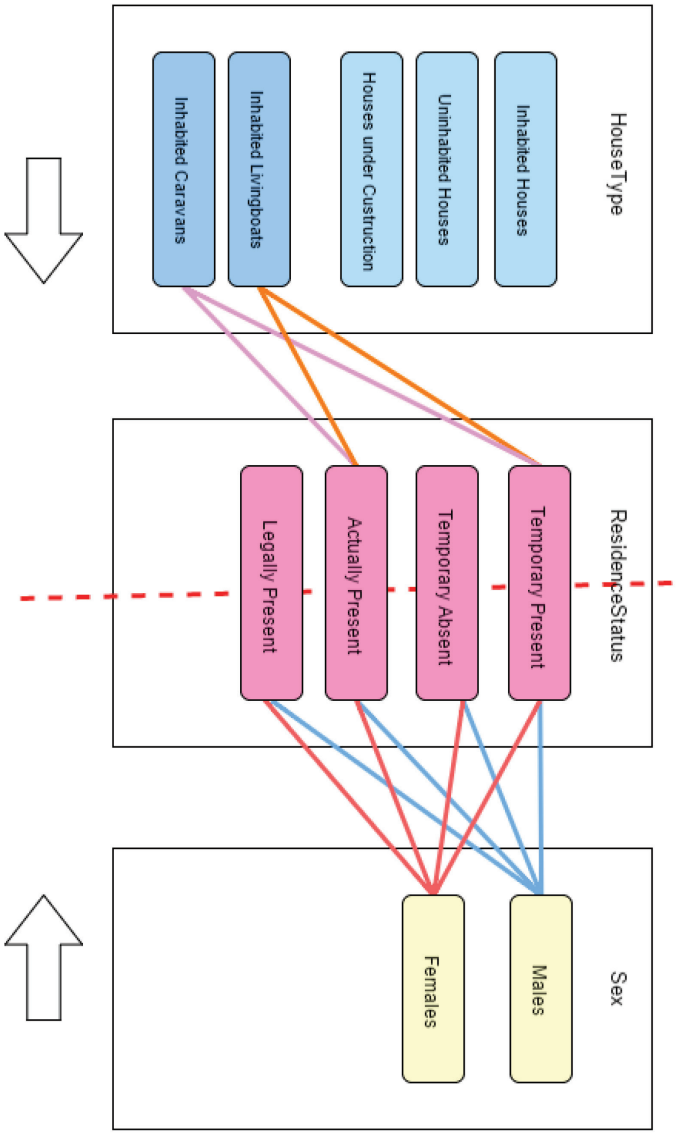


Figure 7.7 - Variable Mappings - Overview of the created Variable Groups, their Values and Mappings

Mappings are important to avoid invalid questions on the data. For example, *without* looking at our standardization model above, users which are interested in a very basic demographical statistic such as *'the total number of males in a certain city'* will get the wrong number back when they simply query for *exactly* that. If we look at Figure 7.7 we see that the variable value *males* is connected to *four* different 'ResidenceStatus' values. So merely asking for the total number of males *without* a 'ResidenceStatus' *restriction* would give us the total number of *males* which are:

- Temporary present
- Temporary absent
- Actually present and
- Legally present

What the user really needs is the combination of the standardized values *males* and *legally present* (to avoid major over counting). In another example, enthusiastic and RDF savvy users will try to make their own queries and ask for variable combinations and questions which are simply not there. For instance users would try to query for *temporary absent ships* which could seem a logical question as we *do* have the values *temporary present* and *ships*. However as we can see from our standardization model the former combination is not possible and would result in an empty or invalid query. Using mappings and documentation about the meaning of the standardized variables and values, users will be able to construct valid queries on the harmonized data and produce sound statistics when querying the data themselves. Therefore, it is the *combination* of standardized variables and values which allow

us to reconstruct the data of the original tables and make *sensible* queries in the RDF database.

Because of our flexible and modular approach we can already start testing our standardized variables and values. After each harmonization change we jump to the '*Testing Stage*' of our workflow and try to improve our standardizations. Because of all the peculiarities contained in the tables, this process is very tedious and requires a flexible approach. The standardization process is one of the most important stages of the harmonization workflow and should therefor get the matching attention. Once the variables and values are given context and standardized uniformly across the years, the process of *creating* variables and *classifying* requires less cleaning and corrections.

THE STANDARDIZATION TEMPLATE

Table 7.4 shows a sample of the harmonization inputs we provide. This example shows the standardizations we have used to define the various values of the variable '*Residence Statuses'* connected with both persons as well as houseboats and wagons in the census.

File name	Literal	Code
cell-VT_1879_01_H1-S0-R5	totaal	JuridischAanwezig
cell-VT_1879_02_H1-S0-R5	totaal.	JuridischAanwezig
	bewoners.	JuridischAanwezig
	bevolking die in de gemeente werkelijke woonplaats heeft.	JuridischAanwezig
	bevolking die in de gemeente werkelijke woonplaats heeft	
	tijdelijk aanwezig	TijdelijkAanwezig
	tijdelijk aanwezig	TijdelijkAanwezig

Table 7.4 - Harmonization Template Format and Input Example for the ResidenceStatus variable.

The Excel files containing the standardizations (such as Table 7.4) are the *input* of the Integrator scripts we described earlier. Our standardization template consists of a three column layout (i.e. filename, literal and code) to connect the *literals* / *labels* contained in the rows and columns of our tables to *standardized* terms, and give exceptions wherever needed. When looking at Table 7.4 we see that the exceptions are recorded in the first column (File name), for example from our mappings we can see that the term Legally Present (Juridisch Aanwezig), is called total (totaal) in

1879. We provide the specific Table and cell locations (e.g. H1-S0-R5) for these exceptions as input for our pipeline and handle these terms accordingly. The Second column of our template contains the literals found in the Excel files which have been extracted in semi-automated ways. Finally, we provide the harmonized definitions provided by expert users in the third column. By following such a structured approach we are able to connect the literals found in the tables to standardized terms (in bulk), and deal with exceptions if needed.

7.2.4 CLASSIFICATION

Once the data has been standardized and tested, we move on to the next stage of our harmonization workflow, i.e. classification of the data values. In this stage, all variables which contain numerous different values, are grouped together into meaningful classes (Begthol 2010). Classification systems come in different forms and from various domains. They often serve specific needs and views of researchers and bring order when working with large amounts of data. In the censuses there are various variables which, unlike the variable Sex, have many possible values. The variable municipality contains around twelve hundred municipalities represented by thousands of literals, there are hundreds of different lower level municipal areas, thousands of literals referring to different religions, thousands of occupations and occupational classes and around three thousand different literals referring to housing types. They all need to be classified.

" It is by the aid of Statistics that law in the social sphere can be ascertained and codified, and certain aspects of the character of God thereby revealed. The study of statistics is thus a religious service."

Florence Nightingale [81]

"Know the past to know the future" is a thought which, if we look at the census, has not been made easy to do. In the case of the Dutch historical population censuses, we have three main classification systems (see 7.2.2 Inspection), i.e. census years which share the same structure and variables. Bridging the gap between the different classification systems, using aggregate data, is not always possible without creating our own classifications *or* variables (see next section).

Our classification approach is a twofold one (see Figure 7.1 with feedback loops to external classifications and variables). First of all, we make use of the advantage of having everything exposed in the Semantic Web which makes it possible to connect to *existing* classifications systems wherever possible. In order to see which variables or classification systems are currently available we have built a tool (LSD Dimensions) to 'scan' the Semantic Web and provide an overview of available classifications and variables. Next to that, we create our own bottom up systems to accommodate the lack of standard variables and classifications in the Semantic Web. As we are aiming to harmonize *historical statistical* data, and a very *specific* one as the *censuses*, we found that the majority of the

[81] Maindonald, J. and Richardson, A. M. (2004) 'This passionate study: a dialogue with Florence Nightingale', *journal of Statistics Education*, vol 12, no 1

variables which we are interested in are just not in the Semantic Web, yet. Except for relatively simple variables such as Sex and Marital Status, which are provided by the *SDMX* (Statistical Data and Metadata eXchange) vocabulary. Consequently, during the initial inspection stage we already realized that almost all of the classification systems and standardizations we were interested in had to be made by ourselves. By creating and providing our classification systems in open formats such as RDF, we are not only harmonizing our own dataset, but also *enriching the web* with our definitions and variables which can be easily reused by others.

What we need are census specific source-oriented harmonizations, starting with a frequency list of all the different values for municipalities, religious denominations, residence statuses, housing types etc. As we have defined the location of the variables during the conversion of the Excel files into RDF, we are able to query all the values of a specific variable to create frequency lists as a basis for semi-manual classifications. The classifications described below are created and based on our harmonization needs, using the expertise of frequent data users. In the following we give examples of these type of bottom up classification and the connections to external systems.

Housing types such as barracks, wagons, ships, institutions, hospitals, monasteries, prisons etc.. are used throughout the population census in different degrees of detail. Whereas in some years we have detailed information such as 'the asylum of Saint Paul' or 'the abbey of Berne' in other years we have only information on the aggregated level of such housing types, i.e. asylum or monastery. The former detailed cases, although

interesting for local historians, would not be of much use for researchers interested in longitudinal analysis. As said, we need to put these (detailed) values into usable groups based on the function they perform (hospital, military buildings, mental institutions etc.). By doing so we have created a bottom up classification system which for the first time allows us to analyze the evolution of the different housing types in the Netherlands over time with marginal effort. From the number of care homes for the elderly, to mental institutions, to the number of forts or barracks, a variety of interesting house types are now standardized and classified using automated frequency lists and expertise of knowledge users in the project. This housing classification resulted in the grouping of over 2000 unique literals (unique terms found in the census tables) into fourteen major classes and thirty-one minor classes. Table 7.5 provides an overview of the defined classes and their associated codes for the housing classification system.

The Dutch Census Housing Classification System

Huizen	1-0
Overige Huizen	2-0
Tijdelijke huizen	2-1
Schepen	2-10
Woonwagens	2-11
Keten	2-12
Medische gebouwen	2-2
Ziekenhuis	2-21
Verplegingshuis	2-22
Gesticht	2-23
Instituut voor Doofstommen	2-24
Herstellingsoord	2-25
Verzorgingshuizen	2-3
Oude mannen en Vrouwenhuis	2-31
Oude vrouwenhuis	2-32
Weduwenhuis	2-33
Rusthuis	2-34
Onderdak voorzieningen	2-4
Gasthuis	2-41
Armhuis	2-42
Weeshuis	2-43
Godshuis	2-44
Doorgangshuis	2-45
Militaire gebouwen	2-5
Militaire faciliteit gebouwen	2-51
Kazerne	2-52

Fort	2-53
Wachthuis	2-54
Scholen educatief	2-6
Scholen	2-61
Kostschool	2-62
Pensionaat	2-63
Religieuze gebouwen	2-7
Liefdegesticht	2-71
Klooster	2-72
Seminarie	2-73
Religieuze Stichting	2-74
Diaconiehuis	2-75
Huis van bewaring	2-8
Gevangenis	2-81
Werkhuis	2-82
Hofje	2-9
Hotel	3-1
Instituten	3-2
Vereningen	3-3

Table 7.5 – The Housing classification built for the Dutch historical censuses

Municipalities are the most used geographical level after provinces in the census. The census is one of the, and sometimes the only systematic, historical sources for researchers providing comprehensive geographic coverage and broad chronologic scope (Ruggles and Mennard 1995). However, boundaries of municipalities may change over time, as well as their names, severely hampering longitudinal studies. Historically the boundaries of the municipalities in the Netherlands underwent major changes. Between 1812 and 2006 there were only six municipalities which did *not* experience changing boundaries (Van der Meer and Boonstra 2006). Besides this, there were a lot of changes in the spelling, see Figure 7.8 for changes in spelling of a municipality at different road sides! In order to track these changes several (external) classifications have been developed (Amsterdam Code, CBS Code, Wageningen Code etc.) to allow comparisons over time and space. We use the AMCO (Amsterdam Code) as the main classification system to harmonize the municipalities in our dataset. Not only does this classification cover the entire time span of our dataset, it is also built on the principle of *minimum varying* codes. In other words, municipalities get fixed codes over time and the system does not alter with changing names, composition or spelling variants (Van der Meer and Boonstra 2006).

Figure 7.8 - Spelling variants of the same municipality at different roadsides taken in 2006 (Van der Meer and Boonstra 2006)

Sub-municipal areas such as districts, neighborhoods and streets have been recorded from 1849 onwards throughout the Dutch historical population censuses. As we showed in Figure 7.6, municipalities are among the minimum required variables which need to be defined in order to get a total for a specific year and place. However, this total is made up of data from the sub-municipal levels. It would be interesting to be able to zoom in on these data. These lower level areas in the historical censuses have been neglected by researchers for comparisons over time, due to consistency challenges. Different years use different levels of detail and ways of organizing the sub-municipal levels, making it difficult to do any type of longitudinal analysis. We build on the work of Boonstra (2007) and identify "Kom" (Inside/Outside the Center) and "Wijk" (District) as the only two sub-municipal variables which are usable for comparisons over time. Localities are unfortunately too inconsistent and not structurally used through the different tables in order to be systematically harmonized. Kom and Wijk two are the most frequent lower level

variables and available for almost the entire range of our harmonized subset of the data, i.e. 1859-1920.

Although the data on *Kom* and *Wijk* level are present, they have been poorly transcribed (Boonstra 2007, Ashkpour, Meroño-Peñuela and Mandemakers 2015) making it difficult to identify and utilize them. By querying the data as a whole and using basic NLP techniques and scripts we are able to identify each and every cell where a certain *wijk* or *kom* occurs and use these frequency lists for bottom up classification purposes. Applying this approach around 90 percent of the "kom" values were identified in such an automated way, using scripts and standardization rules provided by experts. The last (missing) 10 percent still needed to be identified, however this was not a manual job. The standardizations provided by expert users did not cover all cases and around 10.000 exceptions of "koms" (i.e. literals) still needed to be identified and standardized after running our scripts on the data. For example, during the testing stages we quickly noticed missing data for the census year of 1859. The problem here was that for the tables of 1859 transcriptions errors were made which resulted in the literal "Kom" being used in one line (string) next to the other variables, instead of having its own column. To deal with this particular example we used specific rules and generic scripts to identify whether a cell contained the term "Binnen" (inside) or "Buiten (outside)" de *Kom* and marked them as exceptions to include the 10.000 missing values for 1859.

The (simplified) diagram and Table below gives an overview of the geographical variables and their hierarchy. Municipalities are the highest geographical variables which are explicitly defined in the tables (see Figure 7.9). They are classified using the AMCO

and can be aggregated to present the data on province and national level. As we can see in this Figure the next levels are Kom and Wijk. The "Kom" variable has the values "inside" or "outside" the "Kom", however they could be connected to many different "Wijken". For example, in Alkmaar the "Kom" is divided into five "Wijken" whereas Amsterdam even has fifty in the census of 1899.

Figure 7.9 - different geographical levels of the historical Censuses

7.2.5 A LEXICAL AND SEMANTIC CLASSIFICATION APPROACH

Next to building classification systems using frequency lists and knowledge intensive work, we have experimented with more automated ways to better assist researchers when classifying large amounts of data. In this section we explore how a lexical and semantic classification approach could help researchers in dealing with large amounts of values which need to be grouped.

The amount of historical classification systems currently available in the Semantic-Web is inadequate. Therefore, standardizing and classifying our data using *existing* classifications (i.e. called a top down approach) would not suffice. The creation of bottom-up (i.e. data driven) classifications is a necessary but often a manual task that requires lots of expert *knowledge* and *time* investment. In contrast to certain variables or classification systems which have a limited number of values, other variables have thousands of values which have to be put in meaningful groups. This is for example the case for the different housing types, occupations or municipalities found in the census.

In this section we explore a method which helps expert users to make classification systems out of thousands of values. By *assisting*, not replacing, the expert in the grouping of thousands of non-standardized literals into meaningful groups, we aim to reduce the 'information deluge' which these users are often confronted with. We propose a highly reproducible method to automatically generate classification systems from non-standardized values in RDF graphs in a bottom-up way to *assist* expert users.

RELATED WORK

When working with non-standardized statistical data, the process of creating classification systems has been a mostly manual job. Current classification practices are therefore based mainly on data-driven, bottom-up, manual efforts by domain experts (Esteve and Sobek 2003). Researchers which lack programming skills, budget or sometimes necessitated by the data itself are bound to use (a combination of) different tools in order to clean, filter, group and classify statistical data before its publication: this is the purpose of the OpenRefine tool. A set of clustering algorithms (defined as "finding groups of different values that might be alternative representations of the same thing[82]") are provided. Perhaps (Knijff et al. 2013) is the closest match to the taxonomical knowledge construction via hierarchical clustering that we aim at, although fundamental differences apply with respect to the input data (i.e. collections of documents instead of flat literal lists) of different domains. Unfortunately, there is hardly any tool support available for conducting this classfification: (a) in a purely Linked Data setting; and (b) standardizing values after their publication as RDF in order to preserve both original and standard values.

BOTTOM-UP CONSTRUCTION OF CLASSIFICATION SYSTEM: ASSISTING THE EXPERT

We propose a workflow to automatically build bottom-up classification systems from flat lists of non-standardized dimension values. The (semi)automated method we propose when classifying variables with many possible values is divided in

[82] https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth

five steps, namely: retrieval of literals, hierarchical clustering, semantic tagging, linking and serializing.

Retrieval of the literals of values contained in the raw data is the first step. Since we are interested in building classification system with data already in RDF, we have developed standard SPARQL queries in the form of templates. Once executed, the result set contains a list of unique non-standard value literals (i.e. a frequency Table) which we build on in the next steps.

Hierarchical Clustering is applied after retrieving the literal values we are interested in. In our approach the role of the expert user is central in creating meaningful harmonizations. Our hypothesis is that knowledge experts group disparate literals mostly on a string similarity basis. Obviously, some literals may be grouped together for other reasons (e.g. semantic similarity), and it is part of our approach to understand which ratio of the target concept scheme can be reached using lexical criteria only. Since concept schemes are taxonomies, we choose hierarchical clustering as our method to build taxonomic relations between non-standard literals. To achieve this, we use the result set of the previous step as input for the hierarchical clustering algorithm which we use in SciPi[83], and the *Levenshtein edit distance* (Levenshtein 1966) as a distance metric.

Semantic Tagging is next used to explore additional grouping of the literals. An important task knowledge experts are assigned with when building classification systems, is to label upper categories. For example the cluster containing "Barracks", "Arsenal" and "Citadel" may be named "Military buildings". We suggest

[83] Open Source Python library used for scientific computing and technical computing

meaningful names for the output clusters of the previous step by using semantic resources like WordNet and DBpedia. Concretely, we offer two alternatives for semantic tagging of clusters.

First we use *'Term-based'* tagging. After the removal of certain stop words, we tokenize and stem all literals under the same cluster and rank them according to their appearance frequency. We use the token with the highest frequency to query WordNet and DBpedia and use it as suggestions to name the cluster. Second we use the *'Bag-of-words'* tagging. After the removal of stop words, we tokenize and stem all literals under the same cluster. We query WordNet and DBpedia using all tokens of all literals of the cluster. We next use the 'skos[84]: broader' relations to find the closest common broader concept of all literals, and we use this concept as *suggestion* to name the cluster.

We consider all the descendant links below a cluster node k to belong to the same cluster if *k* is the first node below the cut threshold t. We use $t = 0.7 * \max(d(k, i))$, where *d(k,i)* is the distance between the node *k* and any other node *i*.

Linking the original non-standard values to the developed classification system is the next step. Here we actually connect the newly classified standardizations to the raw data. Since we have preserved the URIs of the original values, issuing links between the two is an almost trivial task in RDF.

Serializing according to current standards is the final step. Once we have produced the classification system and the links back to

[84] W3C recommendation designed for representation classification schemes in RDF. http://www.w3.org/TR/2009/REC-skos-reference-20090818/ (2009)

the original values, we serialize both datasets using SKOS and RDF Data Cube, producing URIs (unique identifiers) for all new concepts.

INPUT DATA: HISTORICAL HOUSING CLASSIFICATION

To test our classification approach for assisting the expert, we use gold standards such as the housing classification (section 8.2.4), as an example to compare the results of the automatic suggestions. Our gold standards are classification schemes developed by knowledge experts on top of the Dutch historical censuses. In this example we look at the gold standard for the *historical housing types*[85], where expert-based input is used for classifying the data in a straightforward approach in which the terms are manually classified according to their functions. We compare the results of the automated classification methods with expert-crafted classifications (e.g. the housing type classification) to compare in which degree they could assist expert users in their endeavors of putting thousands of literals into meaningful groups.

EXPERIMENT RESULTS

In order to explore and find the appropriate parameter values for hierarchical clustering we execute our classification approach several times using different parameters. We take the average term distance to determine the distance between clusters. Figure 7.10 shows our resulting schemes.

[85] See http://goo.gl/Hsqwz0 for the input house types

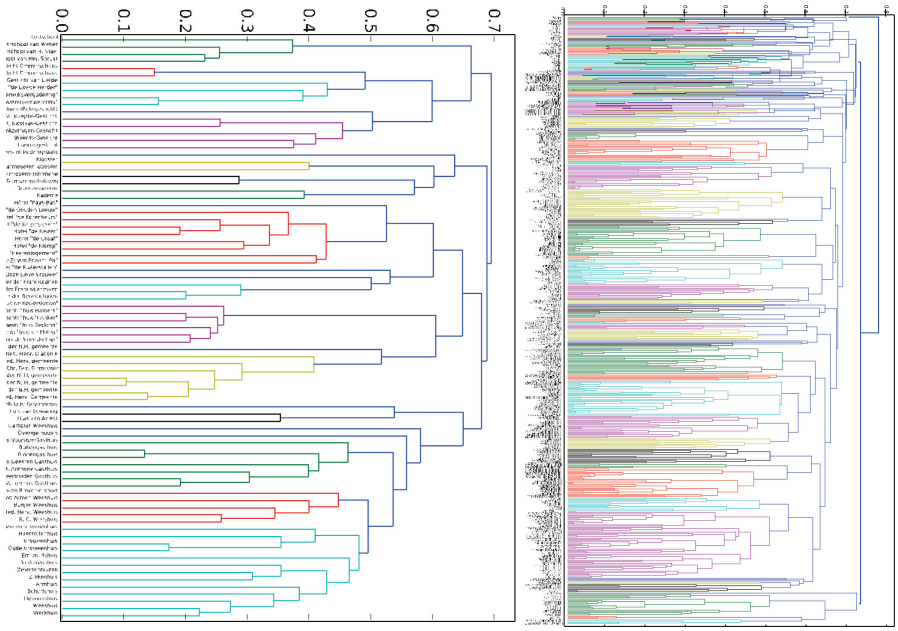


Figure 7.10 - Dendograms of the hierarchical clusters suggested by this approach for the Housing Types

In this example we observe interesting groups being identified in the 'housing types' dataset. For in-stance, the cluster containing the values "Klooster der Franciscanen", "Klooster van de orde der Franciscanessen" and "Klooster van de orde der Benedictijnen" clearly identifies kloosters (monasteries), and gets appropriately http://nl.dbpedia.org/resource/Klooster (Monastery) as a semantic tag for the broader category of the concept scheme. Interestingly, a purely lexical approach exploit the transitivity of some string similarities. For example "Kazerne" and "Militair Ziekenhuis" are clustered together due to the linking member "Militair Kazerne" of the same cluster. On the other hand, the purely lexical clustering also shows its limitations when instances like "ziekenhuis" (hospital), "armhuis" (poorhouse) or "weeshuis"

(orphanage) are clustered together (due their common suffix "-huis") despite their notable semantic differences.

Knowledge experts (i.e. the creators of the 'manual' housing type classification system) validating our workflow compare these results with the gold standards they have set, and see its usefulness when building classification. Concretely, they are interested in its application as a knowledge support tool in the classification system building process. Accordingly, a key issue of the process, covered by our workflow, is using the combination of lexical and semantic structuring. Experts truly think that a combination of both approaches is what indeed goes on when they execute the process manually. Our proposed workflow and tool are not meant to be seen as a totally autonomous tool. What makes this method a first step and worth exploring, is that it allows us to provide classification *suggestions* to the experts when dealing with huge amounts of data which have already been translated into RDF (a language which is not meant for humans to read).

CONCLUSION AND FUTURE WORK

In this section we explored an automatic approach to generate classification systems from non-standard values using data which has already been converted to RDF. We propose a method that combines hierarchical clustering to leverage lexical relatedness, with the enrichment from external knowledge bases to leverage semantic relatedness. We systematically compare our workflow output with the gold standards, in order to get precision scores that evaluate our approach. As a result, we produce classification systems that *guide* knowledge experts when classifying the data manually.

Future work on this topic can be extended to 1. Testing this method on other (comparable) datasets, 2. Finding optimal values of the t threshold, set by empirical exploration and 3. Generalizing our method by implementing additional clustering algorithms and more semantic methods for cluster tagging.

7.2.6 VARIABLE / VALUE CREATION

One of the final stages of the harmonization workflow is the 'variable/value creation' stage. The main imperative of this stage is based on the need for *bridging* and *filling* gaps in the final dataset. By *bridging* we create new variables to make comparisons possible across the tables and over the years; by *filling* we create solutions for value gaps in our data. During the previous stages of the workflow we did not apply any harmonization where the *numbers* are actually affected, the focus of this stage however is *exactly* that. In the case of harmonizing micro-data these are typically unnecessary steps, since there is always the possibility to (re)create variables and values according to one's needs. We have created this stage of the workflow in order to bridge between the different census years and compensate the lack of micro data by creating our own variables and values.

Bridging is done because we are interested in *creating* new variables, to make data comparable over the years, or to make implicit data explicit. We identify different types of variable creations. First we create variables from implicit data contained in other variables. Examples of these are the creation of variables for totals of *provinces*, *population* and the creation of values such as *temporary absent* from the *ResidenceStatus* variable. *Provinces* are

not always explicitly defined in the tables but can be constructed by summing the values of the municipalities. The *Population Total* can be constructed by adding up the total number of *females* and *males*. The *value* 'Temporary Absent' from the *ResidenceStatus* variable, was only provided for certain census years. By looking at the difference between the *Legal* and *Actual Population* size, we are able to provide an estimation of the number of 'Temporary Absent' individuals for years where there is no explicit data. However, in the case of dealing with different *age groups* over the years or changing *occupational classes* we have to use statistical computations to create new variables and values which *cannot* be derived from the census. This can be done by using various statistical techniques such as aggregation, estimation, interpolation etc. when required. For example, age groups can be regrouped to make e.g. 11-16, 17-22, 23-28 comparable with 11-16, 17-28 (by adding 17-22 with 23-28) or 11-18, 19-28 (by interpolating the group of 17-22). We *flag* the newly created variables all as *interpretations* in our dataset. The flag indicates what the change encompasses, tracing the harmonized data back to the original sources.

Filling gaps refers to creating *values (numbers)* which are missing in the harmonized RDF database, due to errors occurring during the conversion or simply because these values are missing in the original tables. Basically there are four reasons for occurring gaps: 1. Data entry mistakes 2. Mistakes in the construction of the styling (TabLinker), used to convert the Excel data into RDF 3. Mistakes in the RDF syntax and 4. Missing data from the original tables. However much we try to harmonize everything and deal with all the peculiarities and exceptions, we will always have some exceptions which do not comply with general rules. These

exceptions result in empty cells or 'holes' throughout our harmonized dataset (see examples in Table 7.6). In order to fill these holes, we do not write specific exception rules or dive into the sources to manually identify a mistake for each and every random exception in the tables. To deal with these exceptions we first apply different rules and scripts to *identify* and *estimate* the missing values. For example we found cases where a given variable, is available for several consequent years, suddenly disappears and then returns. Or in other cases, we have data for six consequent years, except the last or first year (see gaps in Table 7.6). We use these *characteristics* as *detection rules and* write generic scripts to identify and fill in the gaps in a separate Table called "GapFiller".

AMCO	1859	1869	1879	1889	1899	1909	1920
10002	335	390	.	442	539	672	.
10071	252	275	283	.	320	364	458
10072	223	273	305	.	.	405	474
10073	209	268	367	378	345	.	470
10035	.	251	314	410	545	654	699

Table 7.6 - Example of produced harmonized Table with an illustration of different types of gaps.

Following this approach, we store these corrections in a separate layer and never make changes to the raw data itself. See Table 7.7 for an illustration of the GapFiller Table, providing different types of corrections to fill in gaps and correct the data. Table 7.8 provides the GapFiller content in the same structured way as in Table 7.6 for illustration purposes.

Census info	Original Value	New Value	Flag Nr
VT_1859_K234-s0	0	195	F2
VT_1879_T147-s0	0	420	F2
VT_1889_H437-s7	0	299	F1
VT_1889_T428-s7	0	342	F2
VT_1899_F317-s0	0	378	F2
VT_1909_01_T_h189	0	397	F2
VT_1920_01_T_h213	0	723	F2

Table 7.7 - Example of corrected or estimated values in the GapFiller Table. *F= Flag... F1 = no value, corrected manually. F2= no value, estimated.

AMCO	1859	1869	1879	1889	1899	1909	1920
10002			420 F2				723F2
10071				299 F1			
10072				342 F2	378 F2		
10073						397 F2	
10035	195F2						

Table 7.8 - Structured Table view of the GapFiller corrections to illustrate the filling of gaps.

The GapFiller Table (Table 7.7) is based on four fields, i.e. 1. the definition of the Table of the census and the cell number 2. the original value (or 0 in case of missing data), 3. the new value and 4. a flag number (description of the type of change according to our flag classification system). GapFiller contains all the corrections (i.e. estimations) which have been spotted by way of

scripts or entered manually by users of the data when they spot mistakes. This file can be used by caretakers of the original data (in our case the archivists at DANS) to improve the raw data and by the software developers to improve the software (e.g. using the exceptions found during testing to build better vocabularies and data linking methods). Using this approach in conjunction with automatic estimation we allow users to improve on the estimated numbers and overall quality of the data.

7.2.7 TESTING

The source-oriented harmonization workflow we propose puts great emphasis on *testing* and positions it as the gateway to the final result, i.e. a harmonized dataset. In our workflow each major data transformation process is directly connected, in an iterative way, to the testing stage. It has to be noted that this is one of the most important stages in the entire process and the most time consuming part. The goal of testing is to eliminate any noise added during the conversion and different stages of standardization, classification and variable creation. This entails that we systematically compare the harmonized output to the original source files in order to make sure that the numbers we produce are correct[86]. By exploiting the structured nature of our Excel tables we are able to test our results using only a part of the data. Once we have tested the results of a harmonized variable for a specific province of a certain year, the other tables (provinces) for that year are also accounted for. This is because the tables

[86] During this process we do not activate the Gapfiller Table to prevent wrong comparisons because of improvement of the original data, GapFiller is used to deal with exceptions found in the final output.

mostly share the same structure per census year for the different provinces. In the following we describe our structured approach in testing the data and present examples of issues which we dealt with during this stage.

Testing entails mainly the construction of longitudinal (SPARQL) queries, using the *standardization, classification and variable creation* outcomes with the *mappings* we assigned earlier. The goal is to produce exactly the same numbers as found in the original Excel tables, but now *harmonized* over the years. To test our data we begin with querying for totals in the tables and use queries which return a single number, e.g. the *total* of *inhabited houses* in *Amsterdam*. In case of suspicious numbers, we use 'detailed' queries producing all the numbers in the Excel tables which made up that specific total for Amsterdam. By doing so we structurally investigate and identify mistakes in the data and our harmonizations which we subsequently improve. Furthermore, an additional (and complementary) way to inspect our harmonizations, is by producing new versions of the Excel files, now ingested[87] with the standardizations we applied earlier. This allows us to map our harmonizations in the original Excel tables. In these enriched tables all the literals (strings) contained in the original column and row headers are enriched with standardized terms provided by us. This allows us to *visually* inspect the mappings every time we encounter a wrong number, by just opening the file and hovering over a cell to see the associated standardizations. In case of suspicious numbers, one of the first

[87] https://github.com/CEDAR-project/DataDump-mini-vt/tree/master/enriched-source

steps is to look if the number we are looking for actually has all the correct mappings and standardizations assigned to it.

TYPICAL MISTAKES

By structurally testing our data after each section of our workflow we have identified several typical mistakes such as; mistakes in the conversion from Excel to RDF, mistakes in the harmonization itself (i.e. wrong standardizations, classifications etc.), issues regarding exceptions, the importance of creating preliminary tables to spot mistakes which otherwise would have been easily overlooked and dealing with software (preservation) related issues. The following subsections give an overview of the most common mistakes we have dealt with and their corresponding solutions:

The conversion: update RDF input

Mistakes in the data could be the result of mistakes in the mapping of the data from the Excel tables into the newly created RDF structure. The conversion of the Excel tables to RDF requires manual input which was defined in so-called stylings. Decisions are made on the basis of the Table layout and knowledge about the data. Poorly styled tables, tables with a specific layout which were not supported (yet) by our tool, forgotten ones or just certain styling choices of which the justification was not easily known beforehand, all resulted in incorrect or missing output in RDF. Some of these mistakes can be spotted directly after converting the data by just looking at the logs, others only when they are compared with the original data. Every time we find a case where a new styling is required we produce new versions which directly

replace the existing ones in our online repository (GitHub)[88] . We refer to this whole process as the Integrator. The CEDAR Integrator[89], is an integration workflow (set of scripts) that automatizes the semantic publication process, i.e. going from original Excel files to RDF ready-to-publish Linked Census Data. The integrator (Figure 6.7) uses the outcome of the workflow to connect our harmonizations to the RDF graph.

Harmonization: update Mappings

The bottom-up approach we follow is one which is coupled with iteration. Harmonization of aggregate historical data should not be a definitive commitment but a learning process. Our flexible approach is built exactly for this. Where we first started with defining a general set of variables, we (at the end of the various iterations) have developed quite specific mappings to deal with the many exceptions and peculiarities which are in the census. This meant that we often had to update our mappings, i.e. add new or correct current standardizations and update the classification codes.

For example, after standardization we directly test and analyze our results. After one of the first runs, we found that we were missing many municipalities. The problem was that we were missing certain combinations for municipalities because of spelling variants (including strange characters). To address this we wrote a repeatable script which produces mappings by setting a certain threshold for the Levenshtein distance, using the standard

[88] https://github.com/CEDAR-project/DataDump-mini-vt/tree/master/source-data
[89] https://github.com/CEDAR-project/Integrator

vocabulary we have built for the tables which *do* work. Once we set a new threshold and ran the mappings we went from 10.000 missing mappings to just 20. To make sure no wrong mappings were applied we manually inspected a sample of the results. The remaining 20 mappings were later coded manually.

Dealing with exceptions

Already during the first step of the harmonization process (standardization) we faced the challenge of dealing with 'context awareness'. In other words, what to do with literals which have multiple meanings? The literal "Huizen" could refer to a municipality in the province of North Holland but it could also simply refer to houses since that is the literal meaning of *huizen,* all in the same Table. In this case we know by expert knowledge that "Huizen" in the column headers *always* mean 'houses' and the ones in the rows are always municipalities. We created RDF queries to extract all the "Huizen" literals and their specific locations in the excel tables to mark them as exceptions. To apply these exceptions we add an extra column next to the original and standardized term in our harmonization input file (see Figure 7.4). In this new column we mentioned the specific location of the exceptions (on three different levels: *Table*, *sheet* or *cell* level) and provide the appropriate standardization for that specific case.

Create preliminary tables

During the first harmonization rounds, we produce many versions of preliminary and intermediate harmonized tables. When the data is still being tested, the 'creating data' stage proves very useful

to identify common mistakes. Having the end result in tables such as Excel or another (relational) Table system, is especially needed when dealing with RDF data. This kind of data are not meant to be visually inspected or read (by humans) and are much less intuitive compared to reading relational databases. The difficulty here is especially that we cannot know what we are missing by looking at the RDF graph database. In order to actually see what we have harmonized and test our result we query the graph database and produce structured tables, to spot certain mistakes in our data.

For example, we know by expert knowledge that the classifications and variables for 1859 and 1869 are quite similar and that there were no major changes in the municipal boundaries during this period. By presenting the data in a tabular and readable form we could clearly spot that the first version we produced had too many changes between those two years, which was unexpected. Upon closer inspection we found that we needed to introduce more standardization variations and add missing cities. Other examples where we clearly saw many gaps in our constructed tables were for the tables of 1909 and 1920. These tables diverge from the rest with regards to how they were transcribed. These tables do not have any clear structural hierarchies, i.e. all variables are contained in one row instead of separate cells and columns (with no clear order, i.e. sometimes separating values with a dot, sometimes with a comma, or in other cases no separator at all). In order to include these years we built custom repeatable scripts to identify all separate values which were contained in one single string based on expert input.

Moreover, using the preliminary tables we just built, we are able to look at outliers and try to correct them. We have built templates to look at the percentage increase between the different census years for each variable and use expert knowledge to identify and correct mistakes.

Processing: update software

We have created an integrated pipeline (i.e. the Integrator) in order to easily connect our mappings to the raw harmonized data. Every time we add new harmonizations, stylings or tables we run our pipeline and produce a new interlinked RDF dataset which is ready for querying. Next to testing the different elements of the harmonization itself, we also acknowledge the need to keep developing and testing our tools, scripts and RDF output. Different scripts and automated processes make sure that our harmonization efforts are translated into RDF DataCube compliant data and made interlinkable. Problems occurring at this stage mostly concern server side issues such as crashes and errors during the conversion process, outdated software resulting in processes not working, versioning of the software, conversion rules (scripts) which need to be changed or improved on by implementing more Semantic Web standards etc. Although rather rare than common these glitches can be prevented by regularly testing and updating the pipeline software. In the long term organizational commitment is needed to maintain the software and make sure the system stays up and running in the future.

7.2.8 CREATE (FINAL) DATASET

At this stage of the workflow we have arrived at extracting the defined variables from the RDF graph and produce harmonized[90] census tables across the years. Once we have followed all the workflow steps several times and are satisfied with the quality of our data we actually make the data available for the scientific community and other end users. We do this in three different ways, putting the user needs at the forefront: Querying the RDF data, creating tables (dumps) which can be downloaded and using a semi-automatic extraction system.

First of all the harmonized data is available for querying via a so-called SPAQL endpoint. To help the users, we provide as many query examples[91] as possible, document it and emphasize how to use the correct mappings. All this aside, we acknowledge that the core users of this dataset (historians, sociologist, demographers but also the public) are not waiting to write SPARQL queries when accessing the data. Therefore next to dissemination via querying we provide the 'harmonized data dumps'. Users would like to have immediate access to the tables by simply having a link to download the data instead of query interfaces. These users have more knowledge of the data itself, are used to working with (big) tables and want to incorporate these files into their *own* workflows and tools with which they are familiar. We create the following *harmonized* dumps:

[91] http://www.censusdata.nl

- Flat Table in *Excel* and *CSV* format (The result of the query output: use this as the input for your workflow and tools)
- Structured Excel Tables (hierarchical harmonized view on the data in Excel format, provides an intuitive overview across years in an eye glance)
- SPSS File (ready to use SPSS File with variables already defined)

We first start with producing the *flat Table* which is the direct result from querying the RDF graph. This flat Table is ideal for researchers to use as an input for their own workflow and tools such as Excel, SPSS or GIS tools. However, this flat Table is not very intuitive for other users to inspect visually. In order to provide a Table which shows the evolution and differences of the variables over time we create *structural tables* similar to the (hierarchical) structure of the original Excel tables. To build these more intuitive tables, we import the flat tables into tools such as SPSS, define the variables and build a structured (hierarchical) Table. It was also this format that we used to (visually) spot mistakes or gaps in the final dataset. Moreover, users who do not want to be bothered by all the intermediary steps in creating their own structured tables and just want to analyze the harmonized data in an eye glance can use these tables to do so.

The third option in our data dissemination focuses on the more general users, which are just interested in looking at specific variables or just want to explore the data without being presented the entire set of variables. To allow this we provide a 'guided variable query' option where users select the variables and values they are interested in and build (valid) tables. In order to allow

this we have created an application (Grlc) which builds on top of existing queries which have been deposited in online repositories (i.e. Github). Using Grlc, any user can now query the data by just selecting the needed variables from a dropdown list and retrieve the numbers they are interested in (Meroño-Peñuela and Hoekstra 2016).

This chapter is accompanied with an interface in the form of a website, www.censusdata.nl, including links to the harmonized data, RDF output, RDF query examples, mappings, documentation, GIS visualizations and more. We aim not only to suggest a workflow or a method but also to show the practical outcomes of the steps we have presented, providing tangible results which are open for all to access (from the images to the harmonized data). We aim to stimulate use of the censuses which up until now were seen more as an 'interesting' dataset rather than a practical research asset. For example, using standard templates users can now simply query the harmonized database and ask for the total number of 'inhabited houses' across the seven harmonized years of our pilot case effortlessly. Prior to our efforts, users had to consult 60 different tables and over 80.000 data cells in the original Excel tables to answer this question and end up spending more time on data integration than analysis.

7.3 ACCOUNTABILITY

Documentation alone is not sufficient to account for the different data transformations. Due to the aggregate nature of the Dutch historical censuses, it is even more important to provide the *trail*

of transformations to the users when harmonizing the data. In order to provide accountability we track and provide the trail of sources (provenance) on two levels. First we account for outcome of our results, and give detailed information on the different harmonization practices applied to make the data accessible over the years. Second we provide the trail to the underlying sources, linking the harmonized outcomes back to the original data, both the Excel tables and the scanned images from the original books.

Provenance of the Harmonized outcomes (the 'Source Trail')

Besides describing our variables, providing valid mappings, documentation etc. we want to account for each and every number we produce in the final harmonized dataset. The software we have developed for the integration pipeline keeps track of *all* the transformations made during the harmonization stages of our workflow. When the data is harmonized we produce different tables and are able to account for each individual harmonized number (a key requirement in historical research). For example, the query '*number of Occupied Houseboats across all the years and municipalities and sublevels'* produces thousands of harmonized results, for which we can provide the complete provenance. We can pick any number from this list and see how this specific value is created and which harmonizations were applied. According to our harmonization results the total number of "Occupied Houseboats", Outside the Center", in "1889" for the municipality of "Achtkarspelen" is 40, see Table 7.9.

Source	Municipality	year	pop
VT_1889_04_H1-S0-K132-h	http://.../amco/10199	1889	3
VT_1889_04_H1-S0-K79-h	http://.../amco/10199	1889	12
VT_1889_04_H1-S0-K11-h	http://.../amco/10199	1889	3
VT_1889_04_H1-S0-K118-h	http://.../amco/10199	1889	1
VT_1889_04_H1-S0-K40-h	http://.../amco/10199	1889	4
VT_1889_04_H1-S0-K56-h	http://.../amco/10199	1889	3
VT_1889_04_H1-S0-K69-h	http://.../amco/10199	1889	4
VT_1889_04_H1-S0-K72-h	http://.../amco/10199	1889	4
VT_1889_04_H1-S0-K101-h	http://.../amco/10199	1889	6

Table 7.9 - Provenance trail of the harmonized outcomes of the number of Occupied Houseboats outside the center of Achtkarspelen, 1899.

The use of source data in social historical research is a given practice. Being able to connect the harmonized outcomes to the original *sources* leaves room open for other researchers to check the original data and make their own interpretations when needed. We believe that this principle should also be applied when harmonizing historical census data.

"I never believe in statistics if I didn't make it myself"

Winston Churchill

As shown in Table 7.9 we are able to trace back the harmonized RDF output to the original Excel files on a cell level. Using standard queries we are able to reconstruct how the total of the variable *pop* (it sums up to a number of 40) is generated, which file(s) and specific cells (e.g K132, K79, K11 etc.) are used to do so. To trace this back even further to the original sources, we make

use of information already contained in the Excel tables and provide the necessary (meta)data to link these to the scanned *images* and *books*, presenting the *year* (e.g. 1889) and *type* of the census (e.g. VT, which stands for the Population Census), the *Table* (e.g. 04_H1), *page* (e.g. 4) and *image number* (e.g. 03-0176). Next to providing the trail of the original sources, we can visualize the entire trail of the *harmonization practices* such as standardizations, classification, mappings etc. which were applied to retrieve this specific number.

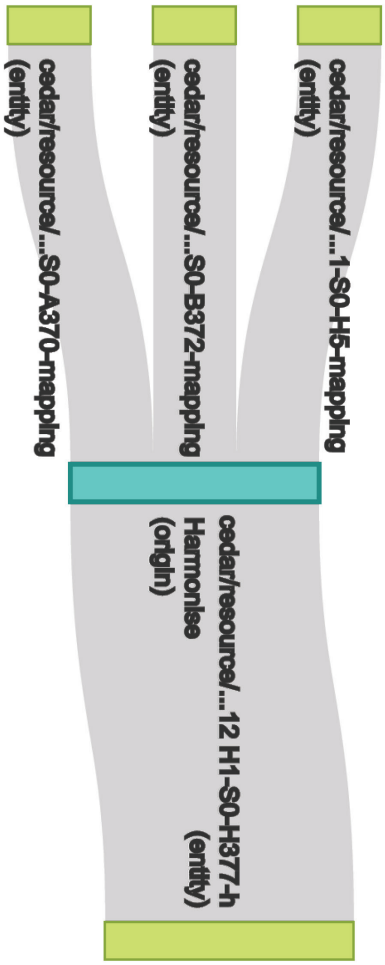


Figure 7.11 - Visualization of the provenance trail

Figure 7.11 shows the visualization of the harmonizations we used for the query "Occupied Houseboats", Outside the Center", in "1889" for the municipality of "Achtkarspelen". For example in this case, the mappings (harmonizations) used for cell H1-S0-H377h were from the S0-H5-mapping, S0-B372-mapping and S0-A370-mapping files. Using this information users can trace back the specific harmonizations and see which standardizations and classifications values were applied in this specific case. For example, the corresponding mappings show that the classification code '10199' is used to harmonize the municipality of Achtkarpselen, the ResidenceType variable for the standardization of the value 'Occupied Houseboats' and the lower geographical value standardized as 'Outside the Center' etc.

By applying provenance at each stage of our workflow we are able to point to the original sources at all times. With this information at hand researchers can consult the original source data and actually see where the data comes from. Moreover, being able to connect the harmonized outcomes to the harmonization practices applied leaves room open for researchers to make their own interpretations when needed.

7.4 STATISTICS ABOUT THE DATA PRODUCED

In this section we show different tables with statistics related to the harmonization inputs we provided by going through the different steps of our source-oriented harmonization workflow. Table 7.10 shows a summary of the different variables and values mapped into RDF observations[92].

Variable	New	Value	# obs
cedar:houseType	X	cedar:house-OccupiedHouses	88,737
		cedar:house-OccupiedShips	28,573
		cedar:house- OccupiedWagons	4,221
		cedar:house-HousesUnder Constructions	14,323
		cedar:house-UnihabitedHouses	51,599
		cedar:house-OtherHousyingTypes	23,344
cedar:isTotal	X	"0" or "1"	205,606
cedar:population	X	xsd:integer	710,462
cedar:residenceStatus	X	residenceStatus : ActuallyPresent	110,668
		residenceStatus : LegallyPresent	220,293
		residenceStatus :TemporaryPresent	119,373
		residenceStatus : TemporaryAbsent	55,733
sdmx-refArea	- /X	From gg:10002 to gg:11447	692,491
sdmx-sex	-	sdmx-code:sex-M	220,661
	-	sdmx-code:sex-F	213,991

Table 7.10 - Number of observations/references connected to the various variables and values harmonized for the Local Division Tables

[92] Observations are the actual occurances of the variables and the numbers in the census tables according to RDF Data Cube

The first column contains the various variables which were defined after harmonizing the Local Division tables. The second column indicates whether we created (X) or reused (-) the variables from existing vocabularies. The third column indicates the available values associated with the variables. The last column indicates how many observations contain such values.

In Data Cube terms observations are the actual numbers in the census the tables. The standardizations we provided for the 'Local Division' tables are together connected to millions of observations in our RDF Graph. By providing expert knowledge and using only the minimum required standardization level we dealt with 2.76 million references. For example, in Table 7.10 we see that the value *OccupiedHouses* of the variable *HouseType* is created by ourselves and linked to 88,737 observations. This means that the standardized term "OccupiedHouses" is connected to 88,737 numbers, i.e. data cells in the census tables. The observations in Table 7.10 are expanded from a number of mapping rules, which we created in the standardization stage of our workflow (see section 7.2.3).

Variable	Mapping file	Generation	#Mappings
City	https://goo.gl/poFcxo	Expert-based/ string similarity	42,294
Housing type	https://goo.gl/fdc0s8	Expert-based	3,484
Province	https://goo.gl/yShX7w	Expert-based	18
Sex	https://goo.gl/ZtVS3z	SDMX	10
Total	https://goo.gl/978YSy	Expert-based SPARQL	38
Housing type situation	https://goo.gl/IEWfBf	Expert-based	22
Residence status	https://goo.gl/TRra0U	Expert-based	40

Table 7.11 - Type and number of mapping rules (standardization synonyms) created per variable type

The first column in Table 7.11 contains the variables concerned. The second column presents the links to the actual mapping files containing the standardizations. The third column indicates how these mapping files were generated: either manually, by purely relying on expert knowledge (expert-based); or semi-automatically, with the aid of querying the raw data (SPARQL) or supported by string similarity scripts. The fourth column indicates the resulting number of mapping rules per file/variable. For example, to harmonize the values of the variable *Housing Type* we have provided 3,484 standardized terms.

Harmonization Results	#Tables	#Cells
Occupied houses and living ships per municipality	60	80,032
Legally registered and present inhabitants per municipality	34	23,086
Houses under construction	47	4,478
Empty houses	60	34,834
Temporarily present inhabitants in ships	35	4,255
Temporarily present inhabitants per municipality	47	74,462
Temporarily absent inhabitants per municipality	34	37,044
Temporarily present inhabitants in wagons	13	426

Table 7.12 – Results of the Harmonized data showing the number of tables and cells

Table 7.12 shows some example queries over the harmonized data produced by going through the workflow we presented in this chapter. For each query / question, we detail the number of tables that users had to open and the number of cells they had to manipulate in order to reach an answer. The data presented in this Table covers the harmonized tables of 1859-1920. SPARQL translations of these queries can be found at http://lod.cedar-project.nl/cedar/data.html and http://www.censusdata.nl. As we can see from this Table some intrinsic major benefits for scholars is efficiency gained in the historical life cycle compared to when using the raw data. For example in the pre-harmonized data, in order to get the *total number of uninhabited houses in the nation from 1859-1920*, users had to consult 60 tables and extract data (i.e. numbers) from 34,834 cells in Excel. And, this assumes that the researchers know exactly where to look for. However in practice researchers end up dwelling around in the census tables and open more than 60 tables.

7.5 CONTRIBUTIONS – THE PERKS OF A SOURCE ORIENTED HARMONIZATION WORKFLOW AND OPEN DATA

After harmonizing the 'local division' tables the data are now open for all to use. The harmonized data is presented in RDF format, stored in an RDF database and repositories for archival purposes. However, as mentioned before RDF is not meant for *humans* to read or to work with efficiently. In order to make the data greater accessible we provide different ways to access and interact with our harmonized data via a web interface (see Figure 7.13).

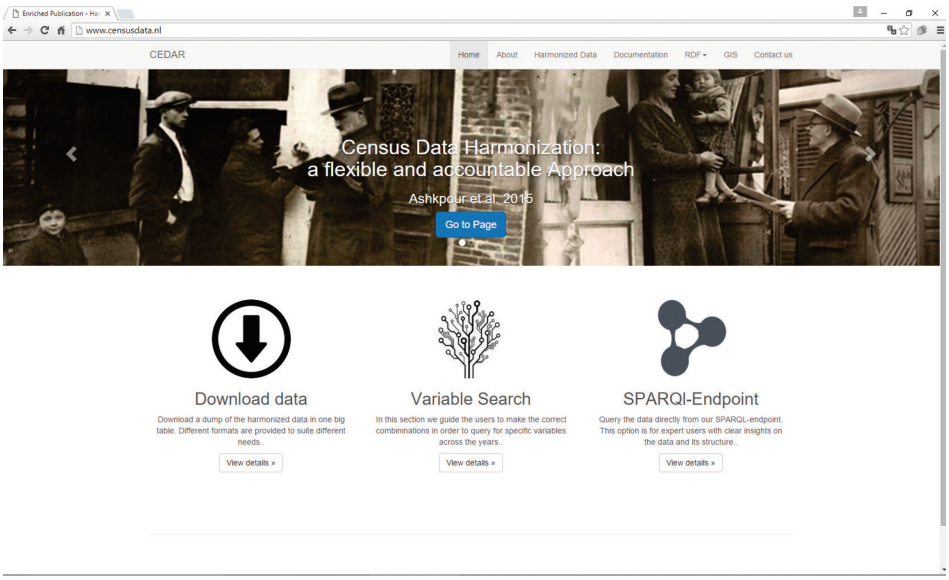


Figure 7.12 - interface designed to access and download the harmonized data in different ways. www.censusdata.nl

The interface in Figure 7.12 is mainly created for the core users of the data such as historians who just want to access to it by downloading the harmonized tables. This interface is built as a 'graphical shelve' around the harmonized census database which is already produced and hosted elsewhere in the CEDAR project (various repositories). This website is a great example of how others could build alternative interfaces and access to datasets of which the data and tools are all *open*. The CEDAR project provides three different ways to interact with and access the RDF data. First, we provide a big harmonized Table via two simple clicks in various formats (CSV, Excel and SPSS) to reduce any extra step to be taken by the researchers themselves (all to stimulate easier use). We provide documentation on the harmonized data and some main publications in the form of articles which describe our harmonization efforts in RDF. Second, next to these tables (data dumps) we also provide a link to the RDF endpoint and various query examples to stimulate users to play around with RDF queries and get more acquainted with how the data can be extracted. See Figure 7.13 for an example query asking for "The total number of inhabited houses, in all municipalities for the period of 1859-1920". Finally, an intermediate step between the dumps and RDF endpoint, is the 'guided variable search' which we aim to provide. Using this option allows users to click through the data and build queries themselves.

1. BewoondeHuizen.sparql

```
# Number of Inhabited Houses per municipality across the years

PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX cedar: <http://lod.cedar-project.nl/vocab/cedar#>
PREFIX sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#>
PREFIX sdmx-code: <http://purl.org/linked-data/sdmx/2009/code#>

SELECT ?municipality ?year (SUM(?pop) AS ?tot)
FROM <urn:graph:cedar-mini:release>
WHERE { ?obs a qb:Observation .
        ?obs cedar:population ?pop .
        ?obs sdmx-dimension:refArea ?municipality .
        ?obs cedar:houseType <http://lod.cedar-project.nl/vocab/cedar#house-BewoondeHuizen> .
        ?slice a qb:Slice.
        ?slice qb:observation ?obs.
        ?slice sdmx-dimension:refPeriod ?year .
        FILTER (NOT EXISTS {?obs cedar:isTotal ?total }) .
        FILTER (?year IN (1859, 1869, 1879, 1889, 1899, 1909, 1920)) .
} GROUP BY ?municipality ?year ORDER BY ?municipality
```

Figure 7.13 - Query Example of the number of 'Bewoonde Huizen' in every municipality across seven census years

LINKS TO OTHER SYSTEMS

Next to creating our own variables (i.e. the various 'Residence Statuses') and classification systems (i.e. for 'Housing Types' or 'Religions'), we already have connected our data with different (external) systems. Currently the CEDAR data is connected to

NLGIS2[93], DBpedia, HISCO, ICONCLASS, Dutch ships and Sailors, AMCO via gemeentegeschiedenis.nl etc. and what has proven to be fruitful in a preliminary stage, see Figure 7.14. In this graphical representation we show the different linked datasets *to* and *from* the CEDAR data.

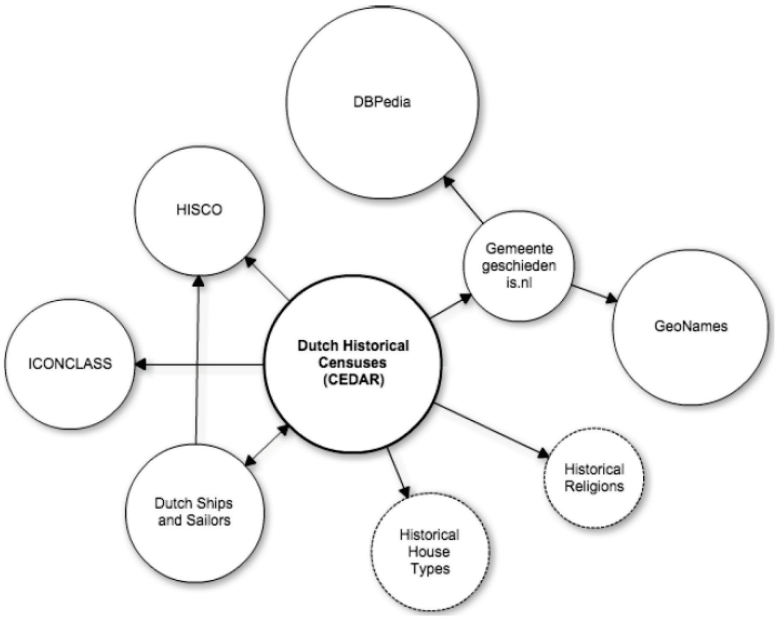


Figure 7.14 - Internal and External Datasets linking to/from CEDAR

Figure 7.14 is a graphical representation of the shared links between CEDAR and various other projects and data sources. The 'links' between these different systems are made by using (common) *standard* Linked Data variables, vocabularies and classification systems.

[93] http://www.nlgis.nl/

Currently we can connect our data to outside sources such as 'gemeentegeschiedenis.nl' to classify our data according to external classification system such as the AMCO for municipalities. Connecting with gemeentegeschiedenis.nl we already have made 2,658,483 links to municipalities in our dataset. Next to this, third parties can also tap into our data and implement it in their own systems *or* provide us their tools, so that we can reuse it on top of our own data. A practical example of this is the NLGIS2 project which uses our harmonized data and visualizes it on historical maps of the Netherlands across time and space (see Figure 7.15). By harmonizing our data and using a GIS we were able to visualize the censuses on historical maps across the various years (i.e. harmonization across time as well as space). Figure 7.15 shows the housing development between 1869-1920. More of these vizualizations can be found via our interface www.censusdata.nl.

In another example, computational musicologists do research with our data by linking the CEDAR dataset with their own historical singers database (Janssen et al. 2015). In these examples, we already gain an easy 'two way' connection by applying Semantic Web standards and its *open* characteristics. In other words, users can tap into our data and simultaneously enrich our data with other sources.

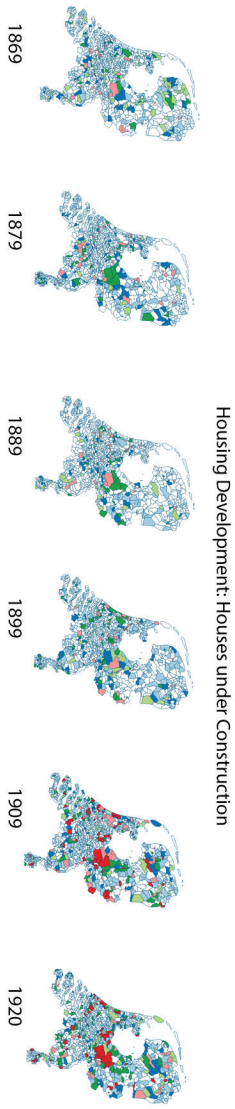


Figure 7.15 - Visualization of the variable 'Houses under Construction' from 1869-1920 on historical maps using NLGIS2. www.censusdata.nl/gis.html

7.6 CONCLUSION

In this chapter we have presented a generic harmonization workflow which builds on the accumulation of knowledge gained in the previous chapters of this study. In order to create this workflow we have looked at the challenges and needs when presented with aggregate data harmonization. In the following sections we looked at the current landscape of census data harmonization and which approaches (the source vs goal-oriented) are best suited to harmonize such data. Using this knowledge we identified a gap between the theory and practice of census data harmonization. We next looked at the application of Semantic Web technologies in historical research and identified how Linked Data technologies such as RDF can help us. Taking this information into consideration we have created an iterative workflow which aims to bridge the gap between the *theory* of how aggregate censuses ideally should be harmonized, and how this is currently happening in the actual *practice*. The source-oriented harmonization workflow we present consist of several stages where each stage is based on the *needs* we have identified in the study (flexible, accountable, bottom-up, iterative etc.). Next to the workflow, we present statistics about the data produced and give practical examples and how to interact with the data.

At this stage we have arrived at a point where we have created generic harmonization methods and applied it on a subset of our data. We also provide the entire Dutch historical censuses in RDF as raw data and build on this to extend our harmonizations. By providing the raw the data next to our harmonized data we aim to stimulate its use by third parties, which we have already seen in the early stages of the harmonization process. Although we believe

that the real reuse of the data will take place on the *harmonized* data. Once the census data is harmonized and made comparable across the years the possibilities are up to the users. This data can be used to create traditional tables for research purposes and to incorporate it into their own workflows and tools, create graphs and other kind of visualizations, connect it to other datasets etc. The key thing here is that the data is harmonized.

By applying the approaches developed in this research we have made several contributions both in practical and methodological ways. To summarize, some concrete outcomes of the harmonization efforts are:

- A raw version of the entire Dutch Historical censuses (1975-1971) is made available as Linked Open Data in RDF (using Semantic Web standards, i.e. Data Cube).
- Creation of historical (bottom up) variables and classification systems which can be extended to other years and similar data
- Linkage with external systems
- A harmonized and highly curated dataset (1859-1920[94]) made available in different formats to accommodate different type of use(r)s.

[94] Currently extended to 2010 by help of a DANS KDP grant

Next to these we have contributed to solutions for:

- Defining harmonization of historical statistical data more concretely
- Providing a structured and flexible and *source-oriented harmonization* method which has proven to be useful for similar datasets
- Providing concrete tools and ways to publish and integrate *historical* statistical datasets in the Semantic Web (a first for historical statistical data)
- Various interfaces and ways to access the data
- Full tracking of our actions, i.e. provenance

As as sidenote to the harmonization of the curated dataset, in 2017 together with the IISH we applied for a DANS KDP (Kleine Data Projecten) grant in order to augment the harmonized census dataset to cover the entire data period of the census and link it with contemporary data. This project ("Linking past and present: augmenting historical municipality characteristics through harmonization and linkage with contemporary data") has been successfully completed[95].

[95] https://doi.org/10.17026/dans-zms-h2s6

8. SUMMARY AND CONCLUSION

The lack of harmonization of the aggregate Dutch historical censuses has been a key constraint when using the data for longitudinal studies. Major changes from one census year to the other and the lack of generic solutions to deal with this prevented many researchers to use the historical census. In this research we worked towards a harmonized aggregate census database and had to overcome many challenges to enable the use of the census in a systematic and longitudinal way. We have combined the theory and practice of harmonization with the principals of source-oriented modeling and introduced an approach that we refer to as source-oriented harmonization. The Resource Description Framework (RDF) has been explored and used as the main technology to integrate the data of the censuses into the Semantic Web. This 'new' Web is an extension of the current Web where information is given well defined meaning and can be read directly by computers.

In this final chapter we start with a summary of the main findings of this study. Next we highlight the most important results and answer the main research question. In the end we present the contributions made and the limitations of this study and provide possible directions for future work on harmonization of historical censuses and other sources based on the lessons learned.

8.1 SUMMARY

8.1.1 HISTORICAL CENSUSES AND HARMONIZATION

In order to understand the challenges of historical census data harmonization, we started the first part of this research with the history of the censuses from ancient times until present. Here we saw a gradual shift in the *use* and *perception* of the census. Changing from a tool from which only 'bad things' could come (as it was mainly used to tax people or for war purposes), to a resource which provides the most comprehensive statistics about nations and fulfilled their information needs. Next, we focused on our specific case, i.e. the Dutch historical censuses and its characteristics. We described the different transformations the Dutch historical censuses underwent throughout the years. The very first transformation of the census already took place in the beginning since the original micro data were not preserved but aggregated in tables and published in the form of books. These books were scanned to preserve the data and eliminate the need for physical access. In a later stage the images (of the books) were consequently transcribed into Excel tables and served as our point of take-off in this study. One of the major challenges we face in this research is directly related to these transformations, i.e. having only aggregate data to work with when harmonizing the data.

We described the importance of having access to the underlying source data at all times, i.e. a practice which we consider one of the most important requirements in the field of historical research. We regard this 'trail of transformations' an important resource for different type of researchers. Next we presented the main problems and challenges of the Dutch historical censuses which

hampered many researchers in using this rich resource. We identified three types of complications. One is the problem of changing variables, values and classifications which is very much related to the changing nature of the census itself. Next to these *changes in content*, are the heterogeneity and inconsistencies in the *structure* of the census tables which also makes it very problematic to efficiently use the historical censuses for longitudinal studies. The final problematic aspects were specifically related to the Dutch census, namely dealing with aggregate data and the need for variable creation. We use these challenges as the bedrock for the harmonization solutions which we address in this study.

In order to find generic solutions for the harmonization of the Dutch historical censuses we studied other census data harmonization projects. We categorized these by looking at key characteristics which influence the harmonization approach itself: source or goal-oriented methods, micro or aggregate data, historical or contemporary censuses and cross year (for a single year) or longitudinal harmonization. The novelty of our approach is explained in this section as we clearly identified a necessity to deal with *harmonization of historical aggregate census data over time*. We found that more current approaches such as RDF are mainly used for contemporary data on micro level and do not harmonize data across time. In general, working with micro data allows researchers to build their own classification systems or create 'new' variables based on their needs. In case of non sensitive historical data one can even go back to the original sources. These are all luxuries which we do not have in the case of the Dutch historical censuses. Interestingly, most of these approaches use RDF only to disseminate the data where the modeling, cleaning, correcting and

standardizations are done prior to converting the data into RDF. In our approach we use a more holistic methodology in RDF which implies that we do all these transformations with the censuses within the framework of the RDF data structure itself.

After describing the Dutch historical census and its challenges we next presented the two main approaches when structuring historical databases, i.e. source vs goal-oriented modeling. The goal-oriented method is often used with pre-defined research questions in mind, when users are interested in a specific part of the data or when users have limited time or budget. Source-oriented approaches however are more inclusive and try to represent the source data as closely as possible. Source-oriented models allow researchers to go back to the original sources and provide different interpretations of the same data. In our study we identified the source-oriented approach as the preferred method of historians. We then urged the neccesity of a flexible and source-oriented harmonization approach when dealing with aggregate historical censuses. Building on the principles of the source-oriented approach and current census harmonization practices, we introduced our own harmonization definition:

" An accountable process of creating an unified and unambiguous version of a dataset, which is flexible enough to deal with the changing characteristics of the data, whilst not committing to a predefined interpretation, by gradually applying a combination of known harmonization practices "

The lack of a clearly defined definition when dealing with the (up until now) ambiguous term 'harmonization', is one of the first problems researchers face when trying to understand what harmonization is and therefore, what its actual practices entail. By

making it more explicit we take away the fuzziness surrounding this term. Using this definition, we show that harmonization is not simply data 'standardization' and data 'cleaning' but that it builds on *a 'set of common practices'*, which sometimes are ignored in current efforts.

8.1.2 HISTORICAL RESEARCH AND THE SEMANTIC WEB

In order to reach our goal of providing harmonized Dutch historical censuses in the Semantic Web, we unavoidably touched upon several interdisciplinary research areas. In the second part of this dissertation, chapter 4 and 5, we look at the fields of Historical Research and the Semantic Web, and the crossroads of these fields, often known as *computing and humanities, history and computing* or *e-humanities*. The work in these two chapters was the result of a comprehensive survey aiming to inventorise the joint work of historians and computer scientists in the use of Semantic (Web) methods and technologies in historical research. We introduced various research efforts in the historical domain (namely papers, projects, online resources and tools) to the field of the Semantic Web and described to what extent historical research can be done using Semantic (Web) technologies.

As the main technology we study and apply for the harmonization of the Dutch historical censuses is based on RDF, we started with introducing the Semantic Web and its principles. We described the Semantic Web as an evolution and extension of the existing 'Web'. The Semantic Web is based on the principles of *structured data*, *meaning* and the use of *standards* in order to better enable computers and people to work in cooperation. Whereas the current Web is currently known as the 'Web of documents', the

Semantic Web is considered the 'Web of data'.

In our survey we took a closer look at historical research and major changes in its methodology, largely due to the introduction of computers and more recently, the Web. We described historical information science and how computer science has inspired historians from its early beginnings. We showed that terms such as 'history and computing' were already being used before the inception of the Web. Since the advent of computing historians have been using it in their research, day to day activities or teachings in one way or the other. Moreover, the use of computers have allowed historians to aim for world-wide, large scale collaborations, especially in the area of economic and social history. With the introduction of Semantic Web technologies in the field of historical research we presented new opportunities for historians to expand these efforts.

Historical information, and the various ways to create, design, enrich, edit, retrieve, analyze and present it with the help of information technology is given specific attention in historical research. Consequently, we presented and followed the *life cycle of historical information* to study the workflow of historians and to analyse which contributions can be made to each of these phases. We noted that the phases, although sequentially presented, do not always need to be passed in a strict order and some can be skipped if necessary. Moreover, the historical sources in this life cycle are traditionally described as primary and secondary sources, however we acknowledged that these are not static notions. Primary sources can become secondary sources for other researchers and vice versa. In order to determine the level of structure of historical sources we

classified different types of historical data according to their level of data structure, i.e. structured, semi structured and unstructured sources.

Next, we discussed typical problems of historical sources, relationships between sources, historical analysis and the way historical sources are presented. We took a closer look at practices of historical research that made use of Semantic technologies. More concretely, we studied contributions with regards to historical ontologies and linking of historical data. Here we classified the contributions in categories of 'scientific papers, research projects, online resources' and 'tools, ontologies and lexical resources'. Next to this we presented historical data integration issues such as dealing with historical ontologies, building links between them and described to what extent relevant contributions consider the problem of data integration and use the Semantic Web to deal with such issues.

In the concluding section we described how the advent of the Semantic Web technologies poses new perspectives, challenges and research opportunities. Historical research is an interesting domain for the Semantic Web as historical data are highly context dependent, have a temporal aspect and are open to a variety of possible interpretations. The concept of a historical Semantic Web however is still a relatively new one and needs to develop more in order to convince historians to change their current methods. We therefore closed the second part of this study with open challenges required for the Semantic Web to further develop with regard to historical research.

8.1.3 HARMONIZATION OF HISTORICAL CENSUSES USING LINKED DATA

In the third part of this study we presented our harmonization solutions for the Dutch historical censuses using Semantic Web Technologies. Almost two decades after the digitization of the volumes with aggregated data of the Dutch historical censues, we are now able to provide generic solutions to deal with the harmonization issues of such data. We based our approach on the theory and practices of historical census data harmonizations *and* on the requirements of historians who are the main users of this data source. In this last section of the dissertation we focused on the creation of a generic harmonization *model* and on a source-oriented harmonization *workflow*.

The aggregated and historical nature of our data required a different harmonization approach compared to the harmonization efforts of contemporary censuses. In chapter 6 we presented a 'three tier' harmonization model in which we have different layers for the *original* (raw) data, the *harmonized* data and the *annotations* (manual corrections made to the data). We showed that this separation is a prerequisite to make the source-oriented approach possible. Moreover by separating these layers, we are also able to provide accountability (i.e. provenance) on two levels. First we are able to connect our harmonizations, on a cell level, back to the original source data. This practically means that we can identify at all stages from which cells of the Excel tables or pages of the books, the modified data are originating. Researchers can therefore use this information to consult the original sources at any time and inspect or modify our harmonizations according to their own interpretations. Second, we provide accountability on

the harmonization itself. We showed how we are able to provide the associated harmonization *rules* at all times and point to the specific standardizations, classification codes, variable creation rules, and estimated values that we find in the final harmonized database. The transparency and flexibility of our approach aims to stimulate others in harmonizing and sharing their (often) confined datasets.

Next, we presented the main advantages of our harmonization methods. The first major difference with current approaches is that we have created a system where we cover the entire process of harmonization in RDF. We have embedded the requirements and workflows of historians into our harmonization approach. We provide methods to deal with cleaning, correcting, harmonizing and data transformations that are related to one another, all in RDF. In fact, we never make changes to the original data, but rather build on top of it in a separate layer. We do this even if obvious mistakes are spotted in the census tables. This permanently allows researchers to consult the original sources upon which our harmonizations are based. This resulted in a unique contribution to research in this field and beyond by presenting a solution to deal with the harmonization of *aggregate* historical statistical data over time in general.

We argued the importance of avoiding an early, concrete conceptualization of the way the data will be used, according to the principles of the historical source-oriented paradigm. We reasoned that a graph data model like RDF proves to be appropriate when the structure of the dataset is very heterogeneous, since RDF is an open structure in which the meaning of the data is defined in the graphs themselves. By

utilizing these benefits of RDF, we can represent the historical censuses with diverse graphs that match their diverse structure, without constraints on meeting an overall agreed schema. Following this approach we provide flexibility in different ways. First, the harmonization workflow we have presented is built in an iterative way. This allows us to *explore* and *learn* from the data, rather than to impose a predefined top down harmonization. Our harmonization workflow therefore greatly emphasizes the importance of flexibility and allowing a learning curve when going through this process. Second, we provide flexibility when *defining* the data. We realize, especially when working with aggregate data, that sometimes it is necessary to create our own variables, including estimated values. These estimations could be interpreted differently depending on the researcher and they may want to provide their own harmonizations. Our system is built in such a way that different standardizations, variable creations or classifications can be applied on the same data. Doing so we fulfilled another key requirement in historical research, i.e. allowing different *interpretations* on historical source data. By doing everything in an open environment (from the tools, scripts, source files, harmonizations, to query examples in RDF), we aim to stimulate greater reuse of these methods, providing generic methods to expose historical data in the Semantic Web.

8.2 RESULTS AND RESEARCH QUESTION

This section presents the most important results of this study and concludes by answering the main research question.

8.2.1 THE DUTCH HISTORICAL CENSUSES CONVERTED INTO THE SEMANTIC WEB

Censuses tend to represent social reality in a very specific way. They are susceptible to change in order to meet the information needs of a specific government or society, providing a contemporaneous view on societal reality. Harmonization of historical censuses is a prerequisite for utilizing the potential of census data for longitudinal research. The aggregated nature of our data and our aim to harmonize the data across time leads us to an approach which is different compared to micro data efforts.

Our harmonization approach builds on the principle that the underlying dataset should be converted into an RDF database while keeping the data structure of the source. By doing so we convert the historical data sources as *one to one* copies in the Semantic Web, as a first step. By converting the data into one system, i.e. RDF, we gain the advantage that we are now able to query the 2,249 census tables as a whole for the first time. This allows us to explore the data, discover its peculiarities and ambiguities. Moreover, we are able to get basic statistics with regard to what we actually have, which was not a trivial task prior to this conversion. We have developed generic scripts and different types of standard queries and interfaces to access and analyze the data, which serves as the input for the experts during the harmonization process. For example, by creating univariate

and hierarchical frequency lists to analyze the landscape of the 'RDF'ed' data, creating query examples, visualizing the raw data for outlier detection methods or on historical maps, producing baseline statistics etc.

In order to realize this we have developed a tool (TabLinker) which converts data from heterogeneous Excel files into RDF data in a very straightforward process. More importantly, we can do this *without* losing any information from the original census tables. RDF has proven to be especially useful as it allows us to build databases using the model, i.e. structure and presentation of the underlying data source itself. In this way we create accurate source-oriented historical databases quite efficiently. By doing so we preserve valuable fine-grained information contained in specific census years for researchers interested in the original categories such as (local) historians and historical demographers. Using RDF, combined with our source-oriented harmonization requirements and solutions was a seamless match. That is to say, a graph data model like RDF proved to be appropriate when datasets suffer from structural heterogeneity. It allows us to convert original source data to a database system without constraints on meeting an overall agreed schema. This was especially true in our case in which we had over two thousand census tables with different structures and level of detail.

To model and harmonize our data we used RDF standards set by institutions such as the World Wide Web Consortium (W3C) and scientific research communities. The changing structure of the census and the ambiguity of the variables required a design which is flexible enough to allow different harmonizations. This is especially true when dealing with aggregate data. To keep track

of the changes and harmonizations we make, we have developed a 'flag system' which takes into account the *original* value of the data, the *interpretations* we assign to the values and the specification of the *actions* which have been undertaken to harmonize or correct the original data. As a result, not only do we provide and deal with *open data* and *tools*, we also make sure that our *practices* are as open as possible.

Currently, we have published the complete digitized historical censuses (1795-1971) in the Semantic Web. With some SPARQL knowledge and help in the form of query templates, users are already able to query the entire census data contained in our raw data layer. By also presenting our raw data online, we allow third party users to build their own datasets, harmonization and/or tools on top of our data. Next to this we provide a highly curated and harmonized dataset of the population census tables for the years 1859-1920. In order to make the data more accessible to researchers with novice or next to none RDF knowledge we have created an interface to suit different user needs. Via www.censusdata.nl users can access the data with minimum (RDF) knowledge required, find numerous query examples and interact with the data, download dumps of the harmonized data, visualize it using GIS and find more information about our approach. Furthermore, links are provided to *all* data used, including the source data, harmonization rules and provenance.

8.2.2 THE NEED FOR A SOURCE-ORIENTED HARMONIZATION WORKFLOW

Harmonization of historical census data, especially in aggregated form and in comparisons over time and space, had been a relatively vaguely defined concept prior to our research. The source-oriented approach is the preeminent and preferred method in historical research, however this is not reflected in current harmonization efforts and projects. In this dissertation we appeal for *more* source-oriented and structured harmonization efforts and provide a workflow to guide researchers in the harmonization process. We claim that the process of harmonizing the data can be made more explicit and generic by following a structured and iterative approach which combines known harmonization practices. Although the challenges, requirements and specific methods of census data harmonization have been thoroughly described in extant literature, the lack of a generic workflow prevented further development and use of the data. In order to make the harmonization itself more reproducible and explicit we have developed a *structured workflow* which builds on the source-oriented paradigm. The workflow we suggest is a generic approach which could also be applied to harmonize other similar datasets, i.e. multidimensional statistical data.

The workflow we have developed is based on the necessity of having a flexible system which allows us to iteratively *explore* the peculiarities of our data. Flexibility is something which is usually not associated with harmonization of historical census data, although different interpretations on the same data is a key aspect in historical research. Our harmonization workflow puts high emphasis on *flexibility*, *accountability* and allowing a *learning curve*

when going through this process. Following a source-oriented approach is especially important in the case of aggregated data since interpreting and harmonizing this kind of data introduces more *ambiguity* compared to the harmonization of micro-data. Our source-oriented harmonization workflow and methods were extensively tested while harmonizing seven consecutive historical census years, spanning from 1859 to 1920. Here we harmonized all the Dutch geographic entities, demographic and housing variables found in the population censuses. This resulted in the creation of a generic workflow, source-oriented harmonization methods, rules and tools which can easily be *extended* to include other years and datasets. The harmonized variables and values on various demographical variables and on geographic areas such as municipalities and quarters, housing types, and residence statuses etc. are *all* present, in some way or the other, in the other census years and can be re-used seamlessly. Because of this, adding additional censuses to the data, is a marginal effort in our system. As a result, the iterative nature of our workflow allows us to easily extend the data with additional years. To test this, we were granted a "Kleine Data Project" (i.e. small data project) of €10.000 by DANS to include and harmonize more demographic variables from the remaining censuses, realizing a timespan from 1795 untill 2010. This enriched dataset is now available via the DANS archive and www.censusdata.nl website.

8.2.3 AN E-HUMANITIES APPROACH AND INTERDISCIPLINARY BENEFITS

This study was conducted within the context of an interdisciplinary research project, i.e. Census Data Research (CEDAR), aiming to advance the Dutch historical censuses for research purposes, using technologies (i.e. RDF) not yet explored to this extent for such datasets. The CEDAR project brought together expertise and research methods from social history and computer science and embraced them as complementary contributions. Applying 'digital' technologies on humanities or social sciences has had a long history going back to the 1950's. Throughout the years such an approach has been referred to in different ways and is currently known as *humanities computing, computing and humanities, e-science, digital humanities, e-humanities* and other variations. Although using different definitions, they all imply the use of digital technologies as complementary methods in an interdisciplinary setting. In the context of the CEDAR project we refer to this as *e-humanities*.

We pursue a novel approach in using RDF and historical data, by looking at *ontological differences* over time and aiming to *harmonize* these differences. These are differences in meaning, context and the relation between the entities presented, e.g. in our case the various classification sytems, variables and values of the census. By creating a structured harmonization model and approach for historical structured data, we also provide a model for other researchers within the humanities to work with and provide clear cross-disciplinary benefits. By way of an open and transparent harmonization approach we aim to trigger a snowball

effect in the adaptation of our approaches with (social) historians and beyond within the field of e-humanities.

The creation of a generic harmonization approach, contributes to traditional historical techniques by providing methods for understanding and handling the seemingly 'invisible' connections between data. The models, tools and methods we have developed in this project enhance scholarship in both the historical and computational domain. The interdisciplinary approach we followed established innovative collaborations around social history and computer science, while *learning* from the challenges coming from humanities research questions. Using computational methods and RDF, historians are able to explore, link and enrich their data in innovative ways, and at the same time make their knowledge-intensive work more easily reusable for others. Computer scientists on the other hand learn from the problems of working with historical data, its life cycle and the challenges it presents, the variation in reseach questions, dealing with the temporal aspect in historical research, importance of source-oriented methods etc. Bringing together expertise and methods from different backgrounds introduced the problem of *finding common ground* and *language* to build on. The CEDAR research project was not exempted from such challenges. We found, especially in the beginning stages of this project, that although a common goal was set for the project as a whole, different views and methods were preferred by the various stakeholders. Dealing with such issues was necessary to advance both research agendas of the different domains and has proven to be a pivotal success factor in an interdisciplinary collaboration such as CEDAR.

8.2.4 MAIN RESEARCH QUESTION

> *" What is the need for historical census data harmonization from a theoretical and practical perspective and how can Linked Data contribute as a new technology."*

The harmonization of *aggregate historical census data* over time and space has fallen behind in terms of practical projects and solutions dealing with its challenges, especially in the case of *longitudinal* studies when compared to micro data projects. Around the time of the initiation of this study in 2012, literature concerning data harmonization of aggregate historical statistical data such as the census, were rather scarce (and are still very limited). Furthermore, there was no clear definition describing the key aspects of census data harmonization. The challenge of this research specifically is that we primarily focused on the harmonization of *aggregate* historical data over time and space, which lacks structured, transparent and repeatable solutions. The fact that prior to our efforts no structural efforts could be identified explains the difficulty when dealing with such data and perhaps the lack of current suitable methods.

In our harmonization endeavors we explored and used RDF as the main modeling technique to harmonize and disseminate the data. Not surprisingly harmonization of dissimilar datasets and data integration issues are also inherent in the paradigm of Linked Data itself. The Semantic Web is currently advocating the (re)use and, more importantly, *linkage* of disperse datasets using Linked Data principles. To quote the World Wide Web Consortium (W3C)

definition[96]:

"..not only does the Semantic Web need access to data, but relationships among data should be made available, too, to create a Web of Data (as opposed to a sheer collection of datasets). This collection of interrelated datasets on the Web can also be referred to as Linked Data"

The realization of such a goal inevitably requires large scale harmonization efforts, on a more diverse set of data than ever before. The interlinking of datasets depends on using *shared* definitions and classification systems to make relationships among the data possible. The advantage of RDF is that we can use it to harmonize datasets which were previously confined in their own realms, while directly contributing towards a historical Semantic Web and enrich it with domain specific knowledge.

In order to answer the main research question of this study we first started with providing a clear definition of the term harmonization itself. Our definition is built on the needs and challenges of harmonizing historical aggregate census data. We identified these challenges by looking at numerous international census data harmonization projects and publications, both using traditional methods as well as RDF approaches. Furthermore, we have looked at the life cycle of historical information and the requirements of historians, to present a definition which can be used for practical solutions.

[96] https://www.w3.org/standards/semanticweb/data

Our harmonization definition introduces several key aspects of harmonization for historical research. The first key aspect emphasizes the importance of a transparent and accountable harmonization process. This refers to the need of being able to link the harmonized output to the original sources upon which they are based at all times. To date, the implementation of harmonization solutions and the decisions made during this process have merely been documented in scientific publications and are often difficult to find and to reconstruct from outside the expert circles. We believe that harmonization (which is highly dependent on expert decisions), especially when dealing with aggregate data, needs to be *open* and as *transparent* as possible.

The second key aspect of historical census data harmonization relates to *source versus goal-oriented modeling*. In our definition we strongly build on the notion of the source-oriented paradigm as the point of take-off. We do so by calling for a *flexible* approach which allows the data to be harmonized in an *iterative* and *bottom up* manner, introducing the notion of 'source-oriented harmonization'. It highlights the importance of allowing different interpretations on the (same) data and more importantly not committing to predefined interpretations when moving from original sources towards harmonized historical databases.

And third and most importantly it defines (aggregate data) harmonization as a *gradual* and *iterative* process which requires a *combination* of known harmonization practices. While data standardization and classification are common and highly used practices with micro data, aggregate data often requires an additional step which we refer to as variable and value creation,

i.e. estimation of missing data. We have combined the different harmonization practices which are applied to historical census data in the form of a structured workflow. Moreover, we have positioned inspection and testing as crucial steps in the harmonization process. Our suggested workflow prevents ad-hoc harmonization of the data and *guides* the users through the different steps in creating harmonized historical databases. This workflow is the practical outcome of our main research question, i.e. the practice of historical census data harmonization. In order to answer our research question we have looked at the discrepancy between the theory and practice of census data harmonization, taking into account the preferred research practices of historians and the life cycle of historical information. We therefore consider aggregate historical census data harmonization across time a *source-oriented, bottom up, iterative, accountable and structured process*, where the role of expert users often is essential.

By creating a structured and accountable harmonization model and approach for historical censuses this study aims to provide a bedrock for other researcher facing similar problems to work with, providing inter-disciplinary benefits. To do so, we have created a workflow to *guide* researchers who are interested in harmonizing historical data across time and space. Moreover, we provide the necessary tools, interfaces and software pipeline in an open environment to accommodate the practical side of harmonization in our study. At www.censusdata.nl we have created an interface for researchers interested in our methods and want to download the harmonized data, visualize it on historical maps over time or simply explore and experiment with the RDF query examples.

8.3 CONTRIBUTIONS MADE

In this study we described the challenges associated with harmonization of aggregated historical census data, identified different harmonization practices and proposed possible solutions in order to deal with problems such as changing classifications, the creation of variables based on aggregated data, structural heterogeneity, visualization of such data and beyond. We found that current harmonization methods lean more towards goal-oriented approaches, making it difficult to go back to the original sources and to allow different interpretations on the same data. Such an approach is an important requirement and practice in historical research. In order to deal with these issues, we have explored the possibilities of source-oriented harmonization of historical censuses over time using generic methods. We looked at similar international projects dealing with such issues and found no generic solutions for harmonizations of aggregate historical data. By making the harmonization process concrete in the form of a structured workflow, we make it easier for others to reproduce our results or apply our methods on similar type of data. Moreover, the final versions of our products: the software we used, our scripts, tools, harmonized tables, harmonization rules, mappings, visualizations etc. are all open and deposited in online repositories and national archives in order to ensure its longevity and to stimulate further use. We provide the harmonized files in different formats, such as Excel and CSV, to avert any intermediary step by the researcher and allow easy direct access. We do this specifically to show the core users of this dataset (i.e. historians) that the results are not bound to RDF output only. Doing so, we aim to inspire more source-oriented harmonization

efforts and *revive* similar datasets in becoming more re-useable for historical research.

As the harmonization of the data is an ongoing process, we have already created generic methods and tools to provide a solution in RDF which is flexible enough to deal with changes and challenges of harmonizing aggregated data, both now and in the future. Interestingly, the CEDAR harmonization efforts are already being followed up and expanded by other projects, i.e. the large scale infrastructural project called CLARIAH or the aforementioned 'Linking Past and Present' project awarded by DANS.

Where prior to our harmonization users had to connect around 80.000 data cells to answer a simple question such as '*what is the number of total inhabited houses across the Netherlands during the period 1859-1920?'*, they can now do it in mere seconds. Although we provide all the benefits of RDF we consciously want to shield (some) users from the RDF output. All results are provided in the form of an integrated RDF database, harmonized tables (in formats such as Excel, CSV, SPSS etc.) and an interface to necessitate the different needs of the users. These could be expert users interested in big data dumps (to integrate the data into their own workflows), novice users interested in browsing and exploring the different variables across time, or computer scientists wanting to query the RDF database directly to build their own queries or applications on top of our data. We have developed standard templates and interfaces for querying the data in a uniform manner and experimented with different visualizations to explore our dataset. Our visualizations mainly served two goals. Namely, when exploring the data we first used vizualizations to find outliers and missing data. However, in a later stage, once the

data were harmonized, we were able to visualize our variables on historical (interactive) maps across time as well as space.

In order to standardize and *link* our data we made use of existing 'external' variables such as sex or marital status defined by SDMX (an initiative which sets standard to accomodate the exchange of statistical data), classification systems such as HISCO for historical occupations and the Amsterdam Code for Dutch municipalities. Next to these we have developed new standards and classifications for variables which were not defined in the Semantic Web yet. We did this for variables, such as residence status, religion, historical housing types and connected these variables and classification systems to our data, while simultaneously enriching the Semantic Web.

With the harmonized data being open for all to access, a 'new' source of data has become available and is open to grasps for all to utilize and answer questions which up until now have been so tedious answer, that they were often avoided by researchers. This study focused mainly on the methodological aspect rather than analysis of the final harmonized results. We leave this final part up to the users.

8.4 LIMITATIONS TO BE ADDRESSED AND LESSONS LEARNED

In our structured harmonization approach we have addressed all the issues that need to be handled to make the Dutch historical censuses comparable across time. We provide different solutions to deal with each of these challenges by structurally going through the different steps of our harmonization workflow in RDF. In this section we present the limitations we encountered when working towards a harmonized census data base, building on Linked Data technologies such as RDF. We finish this section with possible directions for future work based on the lessons learned.

8.4.1 LACK OF HISTORICAL VARIABLES AND CLASSIFICATION SYSTEMS

The availability of variables and classifications in the Semantic Web for historical research is still scarce. Unfortunately, when working with historical data and aiming to use Semantic Web technologies, researchers often tend to create variables and classification systems themselves rather than using existing ones, which after all is the big promise and incentive of using Linked Data. Currently when working with RDF we have to put in more effort than expected, but once there, users can truly start to benefit from the possibilities provided by Linked Data principles. In this project we have created standards for the various housing types, geographical areas, residence statuses etc. which had not been available.

We contribute to the advancement of a *historical* Semantic Web by providing a source-oriented harmonization method which produces variables, values and classifications systems which are specific for historical research. As made evident during our harmonization of the 'Local Division' tables, five out of the six variables we used for the CEDAR workflow exercise were created by ourselves. By harmonizing the historical censuses using RDF, we automatically start contributing towards the reuse of variables and vocabularies in an open environment. The CEDAR project is the first to provide harmonized longitudinal historical census data in the Semantic Web. It is also the first effort of such a large scale socio economic dataset using RDF as the main publishing and modeling technique. As we have shown in chapters 5 and 6, the use of Semantic Web in historical research is being advocated and slowly but steadily gaining momentum in different historical research areas. A continuation of this development provides great promises for researchers to share and reuse their data, within and outside the boundaries of their respective fields. We are already seeing new projects aiming to expand upon our approach by providing similar methods for social and economic historical datasets in general. We believe that convincing historians who have been used to proven methods and tools requires more similar approaches in order to show the true benefits of Linked Data technologies. Given current developments we expect the availability and variety of domain specific variables in historical research to increase, fulfilling the true promises so often advertised by Linked Data.

8.4.2 CUMBERSOME WAYS TO INTERACT WITH THE DATA

Handling and exploring the RDF data is more cumbersome compared to the practice of relational databases or spreadsheet and tables as in statistical packages where users have a variety of tools to view and explore the data. The practices which the users of the data (historians) are accustomed to, need to change considerably when working with RDF and Semantic Web technologies. In our harmonization approach we dedicate different steps in the workflow to interact with the data and provide insights into the quality of the data. We address this issue by writing repeatable scripts and queries which can be run on the *raw* data in order to show what is actually in there (not a trivial task when working with RDF). For example, we have written queries to extract from the raw data the hundred most occurring values per variable, hierarchical frequency lists in order to inspect the relationships and dependencies between the variables, error detecting queries (e.g. find numbers which are not integers or negative values), visualizations for outlier detection etc. Providing and storing SPARQL queries as examples for our users was the closest thing available to access and explore the data with relatively less effort. However, these queries are obviously specific to our data. RDF is still a relatively new concept. What we need is further development of tools and interfaces which allow easier interaction with the data after it has been converted to RDF, i.e. visually browsing and inspecting the raw data, without having to write SPARQL queries. In contrast with relational databases we cannot easily click and browse through the graph data, e.g. to see how a specific variable is referred to in a certain year or visually edit the database content. Although some efforts are being made to solve

these issues, current interfaces and solutions are too immature or technical for historians to work with.

8.4.3 COMPLICATED WAYS TO ACCESS THE DATA

Although somewhat related to the aforementioned limitation, 'access the data' refers to ways in which the RDF data is currently *disseminated* and published. In order to get the harmonized data out of the database two approaches are available in projects dealing with RDF data. The most used and basic way of accessing the data is via the so called 'SPARQL endpoints' where users type in their queries, run it and are presented with the data. But, this requires sufficient SPARQL knowledge which historians usually do not have. Next, even with sufficient knowledge of SPARQL querying, users need to know *what* they can query and which combinations are possible. This last part often seems to be the trickiest part. As we have explained in the mappings section of the standardization chapter, it is the correct combination of variables and values that allows us to create valid queries on the data. Accessing the harmonized data by querying the database therefore is mostly useful for expert users who know exactly what to look for *and* have sufficient RDF knowledge. The second approach aims to serve the core users of the data. Researchers such as historians and historical demographers prefer immediate access to the data. To address this we extract the harmonized data ourselves out of the RDF graph and build 'big harmonized tables' in formats which can easily be used (e.g. Excel, CSV, SPSS etc.). These extracted dumps always represent the latest changes to the data. We do all this to remove the extra hurdle and steps for researchers who want direct access to the data instead of querying the RDF graph.

In addition to these two standard approaches we believe it is necessary to provide all the corresponding queries online for others to reuse and get a better grasp of how these harmonized tables are actually created. We aim to encourage and introduce users such as historians to play around with RDF queries and modify them. What is still missing are mature interfaces which allow novice users to create their own queries by simply selecting the needed variables. In this process the users should be guided, meaning that such an interface shows the options in which the variables may be combined. Future work with regard to dissemination therefore should build on more intuitive ways to extract the data from complex RDF graphs which, as we know, are not meant to be read by humans.

8.4.4 RD... WHAT ?!

When it comes to RDF and the Semantic Web there is still much knowledge to be gained in the historical community. While touching upon similar subject matters and methodologies in their daily practices, the possibilies and uses of RDF are not yet utilized effectively by historians. However, historians are no strangers with large historical databases, the use of digital methods, semantics or even Web technologies. As the practical applications of Linked Data technologies grows and matures we expect greater participation and contribution by historians.

Similar to the acceptance of digital methods, the acceptance of RDF within the historical research community is not something we expect to happen overnight. Even ten years after the introduction of RDF to the scientific research community, Linked

Data and the principles of the Semantic Web are still relatively new concepts to most researchers when it comes to its practical uses and benefits. Computational methods such as relational databases applied in the field of history took quite some time before they were embedded so strongly in the workflows of historians. We expect a similar discourse when it comes to RDF. As the application of Semantic Web technologies for historical research are mostly unknown to historians, therefore also the methods, possibilities and uses of it are currently mostly unexplored.

In order to make RDF truly a success within the historical community, what we need are more similar and structured approaches. Currently historical data are being published in the Semantic Web with the promise and anticipation of easy linkage to other datasets. Unfortunately in practice this often means that the data is only made available in RDF format with no curation or harmonization in mind. The drawback with these approaches is that they contravene their intended goal. Instead of providing greater access to the data it gets even more difficult to use for historians and other core users, because it is the same data but now in RDF. In these approaches often the harmonization and ease of access to the data are absent. In the CEDAR project we provide a structured approach for historical aggregate (census) data harmonization. We show the uses of our approach by providing a highly curated and harmonized subset of the data, interfaces to it, query examples and a practical application to datasets besides the census.

Currently more and more historical datasets are being published and linked in the Semantic Web. As more historical data are being

published and made available online as RDF we foresee a snowball effect where users simply 'plug' their data into to Semantic Web, which in the ideal future, will provide a detailed and wide variety of variables and vocabularies to select from. Greater availability will increase the use and thus introduce more and more historians to the principles of RDF.

8.4.5 TOO DEPENDENT ON EXPERT KNOWLEDGE

Currently, the process of harmonizing raw RDF data relies *too* much on expert input. The availability of standard variables in the Semantic Web for (social) historical research is still quite limited, especially when it comes to variables which have a wide variety of values. Examples of these variables are 'occupations', 'religions', geographical areas such as 'municipalities' or 'housing types' with hundreds of different values which need to be structured into meaningful groups. These values are in turn represented by thousands of spelling variants in historical sources such as the census which all need to be classified. In the classification section of this study we described the different (semi) manually created classifications, needed for the harmonization of the Dutch historical censuses. We next explored how we can assist (not replace) researchers in this knowledge-intensive process. We looked at an approach which considers the lexical and (to some extent) semantic properties of the values in order to group them. These groupings are used by experts to define the different classes. Expert users will always play a key role during the harmonization process. However, we call for more tools and methods to *assist* these users when dealing with RDF data. Efforts taken by for

example tools such as SPSS, OpenRefine or even Excel, which provide a number of very useful clustering methods and algorithms, should be extended to the Semantic Web so they can be applied to data which is already in RDF. We consider the process of *guiding* the domain experts in their endeavors when exploring the possibilities in RDF as a key factor for success and acceptance by historians.

8.5 CONCLUDING REMARKS

Using Semantic Web technologies such as RDF to historical research and the broader field of computational humanities is not a novelty. New in this research is the use of RDF as the data system in which we model, harmonize, publish and query the historical censuses. The harmonization method we presented is based on the needs and practices of historians in order to make an eventual transition from their current, deeply rooted, practices to RDF more seamless.

For the harmonization of the Dutch census we have created generic methods and provided solutions to deal with the problem that a *historical* Semantic Web was almost non existent. Introducing Linked Data (RDF) and Semantic Web technologies to such a challenging historical dataset like the census was not self-evident. Developing new methods and building on new technologies to solve the problems of historical censuses and introducing the Linked Data paradigm to the core users of the data has had its own learning and acceptance curve. For decades, researchers such as historians have been using proven technologies and tools to solve longstanding problems. These tools have been

deeply embedded into their datasets and workflows and are currently still the preferred solution for many. However, exploring new ways and technologies has been the driving force for progress in science and beyond. After all, the interplay between research, development and rising technologies has always proven to be a *contributing* factor rather than a limiting one. In this research we have explored the opportunities and possibilities, which have arisen from the emerging practices of applying Semantic Web technologies on historical data. In our approach we strongly build on the expertise of historians and computer scientists and pursue a common goal. That is, advancing current methods and providing greater access to socio-historical data in this joint venture, often referred to as the *digital humanities* approach. To accommodate this, we claim that Semantic Web solutions need specific requirements in order to be correctly deployed in history. They need to be applied to historical data in a complex, layered and properly adapted context. Good practices and standards and their relationship with the life cycle of historical information, are still needed for the field to continue evolving.

Looking back at the beginning of this study, we started with thousands of heterogeneous census tables, containing almost every thinkable difficulty which a source dataset could possibly have. We used the Dutch historical censuses as a starting point to create generic and integrated methods for comparing structured historical sources and to contribute to scholarly practices with regards to traditional historical techniques. Our contributions are directed at the domains of (social) history, demographical studies, historical information sciences and beyond. With our harmonization approach we filled a gap in the current landscape

of harmonization practices as no solutions were yet available when dealing with aggregate census data. By providing generic harmonization solutions for the historical censuses, we expect researchers to make greater use of the censuses again, now with its full potential, for their own research. We aim to stimulate the use of the census by all others interested in exploring the data and learning about lives in the past. We do this while not keeping the data in a self-contained environment, to stimulate use and inspire new links to the census.

APPENDIX

Paper title	Knowledge modelling	Text processing & mining	Search & Retrieval	Semantic interoperability
Hacking History via Event Extraction	◦	✓	✓	
Exploiting Semantic Web Technologies for Intelligent Access to Historical Documents	◦	✓	✓	
Historical Ontologies	✓			◦
Virtual Knowledge in Family History. Visionary Technologies, Dreams and Research Agendas				✓
Past, present and future of historical information science	◦	◦	◦	✓
Proposed category system for 1960-2000 Census occupations	◦			✓
The Comparability of Occupations and the Generation of Income Scores	◦			✓
Challenges and Methods of International Census Harmonization	◦			✓
Making Sense of Census Responses: coding complex variables in the 1920 PUMS	◦			✓
Semantic Networks and Historical Knowledge Management: Introducing New Methods of Computer-based Research	✓	◦		
Queries in Context: Access to Digitized Historic Documents in a Collaboratory of the Humanities		◦	✓	
Converting a Historical Architecture Encyclopedia into a Semantic Knowledge Base	◦	✓	✓	
Historical documents as monuments and as sources		◦	✓	
Digital Hermeneutics: Agora and the Online Understanding of Cultural Heritage	◦	✓	✓	
Visualizing an Historical Semantic Web with Heml	✓		✓	
Exploring Historical RDF with Heml	✓		✓	
LODifier: Generating Linked Data from Unstructured Text	◦	✓		
CLIO - A Databank Oriented System for Historians	✓	◦	✓	◦
CensSys - A system for analyzing census-type data	◦		◦	✓
A discursive analysis of itineraries in an historical and regional corpus of travels: syntax, semantics, and pragmatics in a unified type theoretical framework	◦	✓	◦	
A Comparison of Knowledge Extraction Tools for the Se- mantic Web	✓	✓		

Table A.1 Reviewed papers. The ✓and ◦ signs indicate a strong and a medium relationship, respectively, between the contributions (rows) and the tasks (columns).

Project name	Knowledge modelling	Text processing & mining	Search & Retrieval	Semantic interoperability
Agora	◦	✓	✓	
BRIDGE	✓		✓	◦
CHoral - access to oral history	✓		✓	◦
Historical Timeline Mining and Extraction (HiTiME)	◦	✓	✓	◦
LINKing System for historical family reconstruction (LINKS)	✓			✓
SCRipt Analysis Tools for the Cultural Heritage (SCRATCH)	◦	✓	✓	
FDR Pearl Harbor Project	✓	✓	✓	✓
North Atlantic Population Project (NAPP)	◦		◦	✓
Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic (CKCC)		✓	◦	◦
Voyage of the Slave Ship Sally	◦	✓	◦	
Multilingual Access to Large Spoken Archives (NSF-ITR/MALACH)	◦	◦	✓	✓
H-BOT	◦	✓	✓	
Clergy of the Church of England Database (CCEd)	✓		✓	◦
Armadillo: Historical Data Mining	✓	✓	✓	✓
Historical Event Markup and Linking (HEML)	✓		✓	
SAILS	✓		✓	✓
CLARIN-Verrijkt Koninkrijk	✓	✓	✓	✓
Historical International Standard Classification of Occupations (HISCO)	✓			✓
Historical Sample of the Netherlands (HSN)	◦		✓	✓
CEDAR	✓	◦	✓	✓
Linking History in Place	◦		✓	✓

Table A.2 Reviewed projects. The ✓and ◦ signs indicate a strong and a medium relationship, respectively, between the contributions (rows) and the tasks (columns).

Resource name	Knowledge modelling	Text processing & mining	Search & Retrieval	Semantic interoperability
Semantic Web approaches in Digital History: an Introduction	✓			✓
Fawcett: A Toolkit to Begin an Historical Semantic Web	✓	✓	✓	✓
Spatial cyber infrastructures, ontologies, and the humanities	✓	✓	✓	✓
SIG Ontologies	✓			✓
CultureSampo - Finnish Culture on the Semantic Web 2.0: Thematic Perspectives for the End-user	✓	✓	✓	✓
Text Mining for Historical Documents: Topics and Papers	◦	✓	◦	
RDF vocabularies for historic place names and relations between them	✓			✓
The Semantic Web for Family History	✓		✓	✓
Data portal for Social Sciences. Open data with SPARQL endpoint	✓		✓	✓

Table A.3 Online Resources. The ✓and ◦ signs indicate a strong and a medium relationship, respectively, between the contributions (rows) and the tasks (columns).

Tool, ontology, lexical resource	Knowledge modelling	Text processing & mining	Search & Retrieval	Semantic interoperability
NLP2RDF	◦	✓		
SIMILE/Timeline	◦		✓	
Gapminder	✓		✓	✓
TokenX	◦	✓		
TAPoR	✓	✓		
SEM event model	✓			◦
OpenCYC	✓			✓
XCES		✓		✓
Dublin Core	✓			◦
GATE		✓		
WordNet	◦	✓		
Framenet	✓	✓		
SUMO	✓			◦
MILO	✓			◦
AskSam		✓		
TEI (Text Encoding Initiative)		✓		✓
SGML		✓		✓
TACT		✓	◦	
Wordcruncher		✓		
Atlas.ti		✓	◦	
NLTK		✓		
FRED	✓	✓		
WAHSP and BILAND	◦	✓		
The Event Ontology	✓			◦
LODE	✓			◦
Semantic MediaWiki (SMW)	✓		✓	✓
Europeana Data Model (EDM)	✓			◦

Table A.4 Tools, ontologies and lexical resources. The ✓and ◦ signs indicate a strong and a medium relationship, respectively, between the contributions (rows) and the tasks (columns).

LITERATURE LIST

Alter, G., Mandemakers, K., and Gutmann, M. P. (2009). Defining and distributing longitudinal historical data in a general way through an intermediate structure. *Historical Social Research*, 34(3), pp. 78–114. Retrieved from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-287180

Anderson, I.G. (2008). History and computing. *Making History*. Institute of Historical Research, University of London.

Antoniou, G., and van Harmelen, F. (2004). A Semantic Web Primer (Cooperative Information Systems). *The MIT Press*. Cambrdige, Massachusetts.

Ashkpour, A., A. Meroño-Peñuela., K. Mandemakers. (2015). The Aggregated Dutch Historical Censuses: Harmonization and RDF. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 48(4), pp. 230-245.

Ashkpour, A., K. Mandemakers and O. Boonstra. (2016). Source Oriented Harmonization of Aggregate Historical Census Data: a flexible and accountable approach in RDF. *Historical Social Research / Historische Sozialforschung*, 41(4), pp. 291-321.

Augenstein, I., S. Padó., and S. Rudolph. (2012). LODifier: Generating Linked Data from Unstructured Text. In E. Simperl, P. Cimiano, A. Polleres, O. Corcho, and V. Presutti (Eds.), *The Semantic Web: Research and Applications. 9th Extended Semantic Web Conference, ESWC 2012, Proceedings,* Vol. 7295, pp. 210–224. Berlin, Heidelberg: Springer-Verlag.

Baader, F., I. Horrocks and U. Sattler. (2005). Description Logics as Ontology Languages for the Semantic Web. In Dieter Hutter and Werner Stepahn (Ed.), *Mechanizing Mathematical*

Reasoning, Vol. 2605, pp. 228–248. Berlin, Heidelberg: Springer-Verlag.

Balakrishnan, S., A. Halevy., B. Harb., H. Lee., J. Madhavan., A. Rostamizadeh., W. Shen., K. Wilder., F. Wu., and C. Yu. (2015). Applying WebTables in Practice. *Proceedings of the biennial Conference on Innovative Data Systems Research*. http://cidrdb.org/cidr2013/cidr2015proceedings.zip.

Beghtol, C. (2010). *Classification Theory*. Encyclopedia of Library and Information Science. pp. 1045-60.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4), pp. 551-572.

Benjamins, J. R. (2004). *A Student's Guide to History*. Boston: Bedfors/St. Martin's.

Berners-Lee, T. (2009). (n.d.). Linked Data - Design Issues.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), pp. 34–43.

Berry, D. M. (Ed.). (2012). *Understanding Digital Humanities*. New York: Palgrave Macmillian.

Bethlehem, J. (2009). *Applied Survey Methods: a Statistical Perspective*. New Jersey: John Wiley & Sons.

Boonstra, O. (2007). *Buurten en wijken in de volkstellingen van de negentiende eeuw. In Twee eeuwen Nederland geteld, by Onno Boonstra, Peter Doorn and Rene van Horik,* pp. 455-470. The Hague: DANS.

Boonstra, O., Breure, L., and Doorn, P. (2004). *Past, present and future of historical information science* (1st ed.). Amsterdam: NIWI-KNAW.

Bos, B., and Welling, G. (1995). The Significance of User-Interfaces for Historical Software. *Proceedings of the Eight International Conference of the Association for History and Computing*, pp. 223–236.

Bouquet, P., Giunchiglia, F., van Harmelen, F., Serafini, L., and Stuckenschmidt, H. (2004). Contextualizing Ontologies. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, *1*(4), pp. 325–343.

Bron, M., Huurnink, B., and de Rijke, M. (2011). Linking Archives Using Document Enrichment and Term Selection. In *Research and Advanced Technology for Digital Libraries. 15th international conference on Theory and practice of digital libraries, proceedings.* (Vol. 6966, pp. 360–371). Berlin, Heidelberg: Springer-Verlag.

Bukhari, A. C., and Baker, C. J. O. (2013). The Canadian health census as linked open data: Towards policy making in public health. Paper presented at the *9th International Conference on Data Integration in the Life Sciences*, Montreal.

Paulos, John Allen. (1998). "Between stories and statistics." *In Once Upon a Number: The Hidden Mathematical Logic of Stories*. New York: Basic Books.

Capadisli, S. (2013). Towards Linked Statistical Data Analysis. In Sarven Capadisli, Franck Cotton, Richard Cyganiak, Armin Haller, Alistair Hamilton, and Raphaël Troncy, editors, *1st International Workshop on Semantic Statistics (Sem- Stats 2013), ISWC*, volume 1549, pp. 61–72. CEUR. http://ceur-ws.org/Vol-1549/article-06.pdf.

Cameron, S., and Richardson, S. (2005). *Using Computers in History*. Palgrave Macmillan

Charniak, E., and McDermott, D. (1985). *Introduction to Artificial Intelligence*. Boston: Addison-Wesley.

Codd, E. F. (1969). *Derivability, Redundancy, and Consistency of Relations Stored in Large Data Banks*. San Jose, California.

Constantopoulos, P., Doerr, M., Theodoridou, M., and Tzobanakis, M. (2009). Historical documents as monuments and as sources. In B. Frischer, J. W. Crawford, and D. Koller (Eds.), *Making History Interactive. Computer Applications and Quantitative Methods in Archaeology (CAA). Proceedings of the 37th International Conference*, volume S2079. Oxford: Archaeopress.

Cook, J. (2003). Primary, secondary and tertiary sources. University Australia. Accessed on: 07-05-2016. http://libguides.jcu.edu.au/primary

Cyganiak, R., Reynolds, D., and Tennison, J. (2014). The RDF Data Cube Vocabulary. Technical report, W3C. Accessed on 08-02-2016 http://www.w3.org/TR/vocab-data-cube/.

Cyganiak, R., Reynolds, D., Tennison, J. (2014) The RDF Data Cube Vocabulary, *World Wide Web Consortium (2012*)

Daniels, R. (2004). *Prisoners without trial: Japanese Americans in World War II*, Hill and Wang Critical Issues.

Den Dulk, K., van Maarseveen, J. (1999). 'The Population Censuses in the Netherlands', in: J.G.S.J. van Maarseveen and M.B.G. Gircour, eds.) *A century of statistics. Counting, accounting and recounting in the Netherlands*, Voorburg/Amsterdam 1999, pp. 303-334.

Denley, P. (1994). Models, Sources and Users: Historical Database Design in the 1990s. *History and Computing*, 6(1), pp. 33-43, ISSN 0957-0144. DOI: http://dx.doi.org/10.3366/hac.1994.6.1.33

DERI. RDF Refine - a Google Refine extension for exporting RDF. Technical report, Digital Enterprise Research Institute, 2015. http://refine.deri.ie/.

Doorn, P. (2012). 'The census and the historical demographer'. p. 30. *Frans van Poppel: a sort of farewell*, Liber Amicorum. Nederlands Interdsciplinair Demografisch Intituut (NIDI). Erik Beekink and Evelien Walhout.

Doorn, P., Jonker, J. and Vreugdenhil, T. (2001). 'Digitalisering van de Nederlandse volkstellingen 1795-1971 : met een nadere beschouwing van de gedigitaliseerde telling van 1899'. *Nederland een eeuw geleden geteld : een terugblik op de samenleving rond 1900*, 41-64. Stichting Beheer IISG.

Doorn, P., and Van Maarseveen, J. (2007). 'Inleiding. Twee eeuwen volkstellingen gedigitaliseerd'. *Twee eeuwen Nederland geteld*, 3-17. DANS – Data Archiving and Networked Services.

Dormans, S., and Kok, J. (2010). An alternative approach to large historical databases. Exploring best practices with collaboratories. *Historical Methods*, *43*(3), pp. 97–107.

Durand, J.D. (1960). The Population Statistics of China, A.D. 2-1953. *Taylor and Francis, Ltd on behalf of the Population Investigation Committee*. 13(3), pp. 209-256.

EEHCM. (2012). Enhancing and Enriching Historic Census Microdata. Accessed on 08-02-2016. www.data-archive.ac.uk/about/projects/eehc.

Esteve, A., and Sobek, M. (2003). Challenges and Methods of International Census Harmonization. *Historical Methods*, *36*(2), pp. 37-41.

Feeney, M., and Ross, S. (1994). Information Technology in Humanities Scholarship British Achievements, Prospects, and Barriers. *Historical Social Research*, *19*(1), pp. 3–59.

Fernández, J. D., Prieto, M. A. M., and Gutiérrez, C. (2011). Publishing open statistical data: The Spanish census. *In Proceedings of the 12th Annual International Digital Government Research Conference: Digital government innovation in challenging times (dg.o '11)*, pp. 20–25. New York: ACM.

Fionda, V and G. Grasso. (2014). Linking Historical Data on the Web. In Matthew Horridge, Marco Rospocher, and Jacco van Ossenbruggen, editors, *Proceedings of the ISWC 2014 Posters and Demos Track, 13th International Semantic Web Conference (ISWC2014), volume 1272.* CEUR-WS. http://ceur-ws.org/Vol-1272/paper_107.pdf.

Fitch, C. A., and Ruggles, S. (2003). Building the National Historical Geographic Information System. *Historical Methods*, 36(1), pp. 41-51.

Flouris, G., Manakanatas, D., Kondylakis, H., Plexousakis, D., and Antoniou, G. (2008). Ontology change: classification and survey. *The Knowledge Engineering Review*, *23*(2), pp. 117–152.

Fox. (1969). Nine step process of historical research.

Gangemi, A. (2013). A Comparison of Knowledge Extraction Tools for the Semantic Web. In *The Semantic Web: Semantics and Big Data. 10th International Conference, ESWC 2013, Proceedings*, Vol. 7882. Berlin, Heidelberg: Springer-Verlag.

Garijo, D., Alper, P., Belhajjame, K., Corcho, O., Gil, Y., and Goble, C. (2014). Common motifs in scientific workflows: An empirical analysis. *Future Generation Computer Systems,* 36, pp. 338-351, https://doi.org/10.1016/j.future.2013.09.018.

Goeken, R., Bryer, M., and Lucas, C. (1999). Making Sense of Census Responses Coding Complex Variables in the 1920 PUMS. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *32*(3), pp. 37–41.

Good, and Scates. (1972). Methods of Research.

Grajalez C.G, Magnello, E., Woods, R. and Champkin, J. (2013). Great moments in statistics. *Significance*, 10(6), pp. 21-28.

Graunt, J. (1977). *Natural and political observations mentioned in a following index and made upon the bills of mortality*. The Johns Hopkins Press

Greenstein, D.I. (1989). A Source-Oriented Approach to History and Computing: The Relational Database. *Historical Social Research*, 14 (51), pp. 9-16.

Groote, P.D. and Tassenaar, P.G. (2007). 'Bevolking en infrastructuur in Groningen en Drenthe, 1820-1915'. *Twee eeuwen Nederland geteld*, 273-298. DANS – Data Archiving and Networked Services.

Gruber, T. R. (1993). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal Human-Computer Studies*, 43, pp. 907–928.

Haslhofer, B., Simon, R., Sanderson, R., and van de Sompel, H. (2011). The Open Annotation Collaboration (OAC) Model. *Computing Research Repository, CoRR*, *abs/1106.5178*.

Haigh, T. (2014). We have never been digital. *Communications of the ACM*. 57 (9), pp. 24-28. https://doi.org/10.1145/2644148

Heath, T., and Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space* (1st ed.). Morgan and Claypool.

Hepps, T. (2015). *The 1900 Census in 1900*. Retrieved 07 22, 2016, from http://b.treelines.com/the-1900-census-in-1900/

Higgs, E. (1996). A clearer sense of the census: Victorian censuses and historical research. *Public Record Office Handbooks*, no. 28. Her Majesty's Stationery Office.

HISSTAT. (2009). Historische Statistieken. Accessed on 08-02-2016. http://www.hisstat.be/hisstat_doelstellingen.php

Hobbs, J. R., Stickel, M. E., Appelt, D. E., and Martin, P. A. (1993). Interpretation as Abduction. *Artificial Intelligence*, *63*(1-2), pp. 69–142.

Horlings, E. (2001). 'Werkgelegenheid en economische modernisering. De structuur van de beroepsbevolking 1807-1909'. *Nederland een eeuw geleden geteld: een terugblik op de samenleving rond 1900*, pp. 65-88. Stichting Beheer IISG.

Hoffart, J, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. (2013). YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence*, 194(28), pp. 3161–3165. http://dx.doi.org/10.1016/j.artint.2012.06.001.

HTK (1980b). Handelingen van de Tweede Kamer 1979–1980, 15800, nr. 91. Rijksbegroting 1980. Hoofdstuk XIII.

Hyvönen, E., Tuominen, J., Kauppinen, T., and Väätäinen, J. (2011). Representing and Utilizing Changing Historical Places as an Ontology Time Series. In R. Jain and A. Sheth (Eds.), *Geospatial Semantics and the Semantic Web: Foundations,*

Algorithms, and Applications, pp. 1–25. Berlin, Heidelberg: Springer-Verlag.

I-CeM. (2014). The Integrated Census Microdata. Unlocking our past. Accessed on 08-02-2016. https://www.essex.ac.uk/history/research/icem

Ide, N., and Woolner, D. (2004). Exploiting Semantic Web technologies for intelligent access to historical documents. *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC)*, pp. 2177–2180.

Ide, N., and Woolner, D. (2007). Historical Ontologies, *chapter Words and Intelligence II: Essays in Honor of Yorick Wilks*, pp 137–152. Springer.

InFuse. (2012a). A free service providing easy access to aggregate data from the UK 2011 and 2001 censuses. Accessed on 08-02-2016. http://infuse.mimas.ac.uk

InFuse. (2012b). A free service providing easy access to aggregate data from the UK 2011 and 2001 censuses. Accessed on 08-02-2016. http://infuse.mimas.ac.uk/help/background.html

Janssen, B., A. Meroño-Peñuela., A. Ashkpour, and C. Guéret (2015). Tracking Down the Habitat of Folk Songs. *eHumanities eMagazine*, 4, http://ehumanities.leasepress.com/emagazine-4/featured-article/tracking-down-the-habitat-of-folk-songs/

Jastrow, M. (1914). The study of religion. *Walter Scot.* London.

Jones, C. (1996). Strategies for managing requirements creep. *Computer*, *29*(6), pp. 92–94. http://doi.org/10.1109/2.507640

Jonge, J. A. de (1966). *Vergelijking van de uitkomsten van de beroepstellingen 1849-1960 (Hilversum)*, dertiende algemene volkstelling 31 mei 1960, dl. 10c

Kalampokis, E, A. Nikolov, P. Haase, R. Cyganiak, A. Stasiewicz, A. Karamanou, M. Zotou, D. Zeginis, E. Tambouris, and K. Tarabanis. (2014). Exploiting Linked Data Cubes with OpenCube Toolkit. *In Matthew Horridge, Marco Rospocher, and Jacco van Ossenbruggen, editors, Proceedings of the ISWC 2014 Posters and Demos Track, 13th International Semantic Web Conference (ISWC2014)*, vol. 1272, pp. 137–140, Riva del Garda, Italy, 2014. CEUR-WS. http://ceur-ws.org/Vol-1272/paper_109.pdf.

Kalus, M. (2007). Semantic Networks and Historical Knowledge Management: Introducing New Methods of Computer-based Research. *Ann Arbor, MI: MPublishing, University of Michigan Library*.

Kaplan, C.P, van Valey, T.L. (1980). *CENSUS '80: Continuing the factfinder tradition*. U.S. Department of Commerce, Bureau of the census.

Katus, J. (1984). Volkstelling in opspraak. Een studie naar de overheidsvoorlichting met betrekking tot de volkstelling van 1971 (proefschrift, Leiden 1984).

KB 1947. Koninklijk Besluit. (1947). In: Staatsblad van 5 februari 1947, betreffende de elfde algemene volkstelling (met daaraan verbonden woningtelling). Staatsblad 1947. nr. H44.

Kemman, M., and Kleppe, M. (2013). PoliMedia - Improving Analyses of Radio, TV and Newspaper Coverage of Political Debates. In T. Aalberg and E. Al (Eds.), *15th international conference on Theory and practice of digital libraries, TPDL 2013, Proceedings.* Vol. 8092, pp. 401–404. Berlin, Heidelberg: Springer-Verlag.

Knijff, J. de., Frasincar, F., Hoogenboom, F. (2013). Domain taxonomy learning from text: The subsumption method versus hierarchical clustering. *Data & Knowledge Engineering*, 83, pp. 54-69

Knippenberg, H. (1992). *De religieuze kaart van Nederland: omvang en geografische spreiding van de godsdienstige gezindten vanaf de Reformatie tot heden*, van Gorcum. Assen.

Knippenberg, H. (2001). 'Polarisatie en versnippering: kerk en godsdienst rond 1900'. *Nederland een eeuw geleden geteld: een terugblik op de samenleving rond 1900*, pp. 65-88. Stichting Beheer IISG.

Kok, J., and Wouters, P. (2013). Virtual Knowledge in Family History: Visionary Technologies, Research Dreams, and Research Agendas. In P. Wouters, A. Beaulieu, A. Scharnhorst, and S. Wyatt (Eds.), *Virtual Knowledge. Experimenting in the Humanities and the Social Sciences*, pp. 219–244. Cambridge, Massachusetts: MIT Press.

Kuczynski, T. (Ed.). (1985). *Wirschaftsgeschichte und Mathematik. Beiträge zur Anwendung mathematischer, insbesondere statistischer Methoden in der wirtschafts- und sozialhistorischen Forschung*. Berlin: Akademie-Verlag.

Kuhrt, A. (1995). The Ancient Near East c. 3000–330 B.C.E. Vol 2. p. 695. London: Routledge.

Lebo, T and J. McCusker. (2012). csv2rdf4lod. Technical report, Tetherless World, RPI, https://github.com/timrdf/csv2rdf4lod-automation/wiki.

Leeuwen, M. H. D. Van, Maas, I., Multinational Interdisciplinary Cooperation, and Miles, A. (2004). Creating a Historical International Standard Classification of Occupations An Exercise in Multinational Interdisciplinary Cooperation. *Historical Methods*, 37(4), pp. 186–197.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), pp. 707-710

M. Hemmje, C. Niederée, and T. Risse (Eds.), *From Integrated Publication and Information Systems to Information and Knowledge Environments*. Vol. 3379, pp. 117–127. Berlin, Heidelberg: Springer-Verlag.

Mandemakers, K., and Boonstra, O. (1995). *De levensloop van de Utrechtse bevolking in de 19e eeuw*. p. 186. Van Gorcum: Assen.

Mandemakers, K., and Dillon, L. (2004). Best Practices with Large Databases on Historical Populations. *Historical Methods*, 37(1), pp. 34–38.

Manso, M.-Á., and Wachowicz, M. (2009). GIS Design: A Review of Current Issues in Interoperability. *Geography Compass*, 3(3), pp. 1105–1124.

Margo Anderson and William Seltzer. (2007). Challenges to the Confidentiality of U.S. Federal Statistics, 1910–1965. *Journal of Official Statistics*, 23, pp. 1–34.

McCaa, R. (2006). IPUMS-International: Integrating and Disseminating High Precision Population Census Samples of the USA, Greece, Europe and the World. *Modern Greek Studies Yearbook*, 22(23), pp. 143–162.

McCatry, W. (2003). *Humanities Computing*. In M. Drake (Ed.), Encyclopedia of Library and Information Science, 2nd ed., pp. 1124–1135. New York: Taylor and Francis.

Meroño-Peñuela, A., Ashkpour, A., Guéret, C., Schlobach, S. (2015b). CEDAR: The Dutch Historical Censuses as Linked Open Data. Semantic Web – Interoperability, Usability, Applicability. IOS press

Meroño-Peñuela, A., Ashkpour, A., Guéret, C., Schlobach, S. (2016). An Ecosystem for Integrating and Web-Enabling Messy Spreadsheet Collections. *Knowledge Based Systems*.

Meroño-Peñuela, A., Ashkpour, A., Rietveld, L., Hoekstra, R., and Schlobach, S. (2012). Linked Humanities Data: The Next Frontier? A Case-study in Historical Census Data. In *Proceedings of the 2nd International Workshop on Linked Science (LISC2012). International Semantic Web Conference (ISWC)* (Vol. 951). CEUR Workshop Proceedings. Retrieved from http://ceur-ws.org/Vol-951/

Meroño-Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., and van Harmelen, F. (2015a). Semantic Technologies for Historical Research: A Survey. Semantic Web – Interoperability, Usability, Applicability, 6(6), pp. 539– 564. IOS Press.

Meroño-Peñuela, A., Guéret, C., Hoekstra, R., and Schlobach, S. (2013). Detecting and Reporting Extensional Concept Drift in Statistical Linked Data. In Proceedings of the 1st International Workshop on Semantic Statistics, ISWC. CEUR Workshop Proceedings

Meroño-Peñuela, A., Hoekstra, R. (2016). "grlc Makes GitHub Taste Like Linked Data APIs". SALAD 2016 — Services and Applications over Linked Data APIs and Data. *ESWC 2016*. Heraklion, Crete, Greece

Meroño-Peñuela, A., Ashkpour, A., & Gueret, C. D. M. (2014). From Flat Lists to Taxonomies: Bottom-up Concept Scheme Generation in Linked Statistical Data. In *Proceedings of the 2nd International Workshop on Semantic Statistics (SemStats 2014), ISWC 2014* CEUR.

Merry, M. (2016). Designing Databases for Historical Research. Accessed on 10-02-2016. http://port.sas.ac.uk/mod/book/view.php?id=75andchapterid=133

Methorst, H.W. (1902). Geschiedenis van de statistiek in het Koninkrijk der Nederlanden. In: Bijdragen tot de Statistiek

van Nederland. Nieuwe volgreeks, XIV. Pp. 152–157. CBS, 's-Gravenhage.

Meyer, P. B., and Osborne A. M. (2005). Proposed Category System for 1960-2000 Census Occupations, *Bureau of Labor Statistics Working Paper 383*.

Moot, R., Prévot, L., and Retoré, C. (2011). A discursive analysis of itineraries in an historical and regional corpus of travels: syntax, semantics, and pragmatics in a unified type theoretical framework. In *Constraints in Discourse 2011*, pp. 14–16.

Morris, T., T. Guidry and M. Magdinie. (2015). OpenRefine: A free, open source, powerful tool for working with messy data. *Technical report, The OpenRefine Development Team*. http://openrefine.org/.

MOSAIC. 2011. Recovering Surviving Census Records to Reconstruct Population, Economic, and Cultural History. Accessed on 08-02-2016. http://www.censusmosaic.org/

Müller, M. (1873). Introduction to the science of religion. *New York: Arno Press.*

Muñoz, E, Aidan Hogan, and Alessandra Mileo. (2013). DRETa: Extracting RDF from Wikitables. In Eva Blomqvist and Tudor Groza, editors. *Proceedings of the ISWC 2013 Posters and Demonstrations Track, a track within the 12th International Semantic Web Conference (ISWC 2013)*. pp. 98–92. Sydney, Australia. CEUR-WS, 2013. http://ceur-ws.org/Vol-1035/iswc2013_demo_23.pdf.

Muurling, S., Mandemakers, K. (2012). MOSAIC Census Inventory of the Netherlands. Final report, (IISG Amsterdam), MOSAIC working paper WP2012-006 (url: http://www.iisg.nl/hsn/documents/mosaic-wp-2012.pdf)

NAPP. (2001a). Project Proposal. The North Atlantic Population Project. Accessed on 08-02-2016. https://www.nappdata.org/napp/proposal.shtml

NAPP. (2001b). What is NAPP. The North Atlantic Population Project. Accessed on 08-02-2016. https://www.nappdata.org/napp/intro.shtml

Nentwich, M. (2003). *Cyberscience: Research in the Age of the Internet*. Vienna: Austrian Academy of Sciences Press.

New World Encyclopedia contributors, 'Numbers, Book of', New World Encyclopedia, 27 January 2015, retrieved from http://bit.ly/2aURWhJ

Nicolaas, J.M.M., and Sprangers, A.H. (2007). Buitenlandse migratie in Nederland, 1795-2006. De invloed op de bevolkingssamenstelling. *Twee eeuwen Nederland geteld*, pp. 19-50. DANS – Data Archiving and Networked Services.

NKJV. (2008). The Chronological Study Bible. New Kings James Version. Thomas Nelson

Noble, P. and Fitch, A. (2006). The National Geographic Information System. Population Association of America Annual Meeting Program. LA, California. March 30 - April 1. Accessed on 08-02-2016 http://paa2006.princeton.edu/papers/61417

Noordam, D.J. (2001). Gezin en huishouden op het breukvlak. *Nederland een eeuw geleden geteld: een terugblik op de samenleving rond 1900*, pp. 89-112. Stichting Beheer IISG.

Oldervoll, J. (1989). CENSSYS: A System for Analyzing Census-Type Data. *Computer Applications in the Historical Sciences: Selected Contributions to the Cologne Computer Conference*, pp. 17–22.

Oudhof, J. and Boelens, A.M.S. (2007). 'Arbeidsdeelname van 50-plus vrouwen 1849-2006'. *Twee eeuwen Nederland geteld*, 207-222. DANS – Data Archiving and Networked Services.

Owens, J. B., Yuan, M., Wachowicz, M., Kantabutra, V., Jr., E. A. C., Ames, D. P., and Gangemi, A. (2009). Visualizing Historical Narratives: Geographically-Integrated History and Dynamics GIS. In *National Endowment for the Humanities workshop. Visualizing the Past: Tools and Techniques for Understanding Historical Processes*.

Parrish, F. Louis. (1941). *The classification of religions: its relation to the history of religions*. Printed by the Herald Press. Scottdale, Pa.

Pasin, M. (2011). Semantic Web approaches in Digital History: an Introduction. http://www.slideshare.net/mpasin/presentations/.

Petrou, I., and Papastefanatos, G. (2014). Publishing Greek Census Data as linked open data. *ERCIM News 2014*:96.

Pineo, P.C., Porter, J., McRoberts, H.A. (1977). The 1971 Census and the Socioeconomic Classification of Occupations. *Canadian Review of Sociology,* 14(1), pp. 91-102.

Putte, B. V. D., and Miles, A. (2005). A Social Classification Scheme for Historical Occupational Data. *Historical Methods*, 38(2), pp. 61-92.

Quetelet, A. (1842), A Treatise on Man and the Development of His Faculties. *Published by William and Robert Chambers, Edinburgh*

Quinlan, N.J. (1990). Domesday Book: Studies. RQ

Riechert, T., Morgenstern, U., Auer, S., Tramp, S., and Martin, M. (2010). Knowledge Engineering for Historians on the Example of the Catalogus Professorum Lipsiensis. In *The Semantic Web -- ISWC 2010. 9th International Semantic Web Conference, Proceedings*. Vol. 6496, pp. 225–240. Berlin, Heidelberg: Springer-Verlag.

Ruggles, S. (2006). Linking Historical Censuses: A New Approach. *History and Computing*, 14, pp. 213-224.

Rula, A, Matteo Palmonari, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, Jens Lehmann, and Lorenz Bühmann (2014). Hybrid Acquisition of Temporal Scopes for RDF Data. In Valentina Presutti, Claudia d'Amato, Fabien Gandon, Mathieu d'Aquin, Steffen Staab, and Anna Tordai, editors*, The Semantic Web: Trends and Challenges. 11th International Conference, ESWC 2014, Proceedings,* vol. 8465 of LNCS, pp. 488–503. Springer-Verlag. Anissaras, Crete, Greece.

Robertson, B, G., (2009b). Fawcett: A Toolkit to Begin an Historical Semantic Web. http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/175/217.

Robertson, B. (2009a). Exploring Historical RDF with Heml. *Digital Humanities Quarterly*, 3(1).

Robertson, B. G. (2006). Visualizing an historical Semantic Web with Heml. In *WWW'06. The 15th International World Wide Web Conference 2006, Proceedings*. pp. 1051–1052. USA: ACM Press. New York, NY,

Ruggles, S., and Menard, R. R. (1995). The Minnesota Historical Census Projects. *Historical Methods*, 28(1), pp. 6–10.

Ruggles, S., King, M. L., Levison, D., McCaa, R., and Sobek, M. (2003). IPUMS-International. Historical Methods, 36(2), pp. 60-65

Scheidel, W. (2009). *Rome and China: comparative perspectives on ancient world empires*. Oxford University Press, p. 2

Schreibman, S., Siemens, R., and Unsworth, J. (Eds.). (2004). *A Companion to Digital Humanities*. Malden, MA: Blackwell Publishing Inc.

Segers, R., van Erp, M., Van der Meij, L., Aroyo, L., van Ossenbruggen, J., Schreiber, G., Jacobs, G. (2011). Hacking History via Event Extraction. In *Proceedings of the sixth international conference on Knowledge capture*. pp. 161–162. ACM Press. New York, NY, USA: http://doi.org/10.1145/1999676.1999705

Shadbolt, N., Hall, W., and Berners-Lee, T. (2006). The Semantic Web Revisited. *IEEE Intelligent Systems*, *21*(3), 96–101.

Sharma, R.K. (2007). *Demography and Population Problems*. Atlantic publishers and distributors (P) LTD

Shvaiko, P., and Euzenat, J. (2013). Ontology Matching: State of the Art and Future Challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1), pp. 158–176.

Sieber, R. E., Wellen, C.C and Jin, Y. (2011) Spatial cyber infrastructures, ontologies, and the humanities. *Proceedings of the National Academy of Sciences of the United States of America*, 108(14), pp. 5504–5509.

Slavakis, K., Georgios B., Giannakis and Gonzalo Mateos. (2014). Modeling and Optimizing for Big Data Analytics. *IEEE Signal Processing Magazine*.

Sobek, M. (1995). The Comparability of Occupations and the Generation of Income Scores. *Historical Methods: A Journal of Quantitative and Interdisciplinary History. Special Issue: The Minnesota Historical Census Project*, 28(1), pp. 47–51.

Sobek, M., and Dillon, L. (1995). Interpreting Work: Classifying Occupations in the Public Use Microdata Samples. *Historical Methods,* 28(1), pp. 70-73.

Sobek, M., Ruggles, S. (1999). The IPUMS Project: An Update. Historical Methods, 36(2), pp. 102-110

Speck, W. (1994). "History and computing: some reflections on the past decade." *History and Computing*, 6(1), pp. 28–32.

St-Hilaire, M., Moldofsky, B., Richard, L., and Beaudry, M. (2007). Geocoding and Mapping Historical Census Data. Historical Methods, 40(2), pp. 76–91.

SweCens. (2011). Encoding and linking Swedish Censuses. Accessed on 08-02-2016. http://www.humfak.umu.se/english/research/project/?code=660

Szołtysek, M. and Gruber, S. (2015). Mosaic: recovering surviving census records and reconstructing the familial history of Europe, The History of the Family, DOI: 10.1080/1081602X.2015.1006655

Tauberer, J., 2007. The 2000 U.S. Census: 1 Billion RDF Triples. Accessed on 12-07-2015: http://www.rdfabout.com/demo/census/

Taylor, L.R. (1933). Quirinius and the Census of Judaea. *The american journal of philology*. 54(2), pp. 120-13. John Hopkins University Press

Thaller, M. (1980). Automation on Parnassus. CLIO - A databank oriented system for historians. *Historical Social Research*, 15.

Thaller, M. (1993). 'What is 'Source-oriented Data Processing?' ; What is a 'Historical Information Science?' ', in: L.I. Borodkin and W. Levermann, Istoriia i comp' iuter. Novye

informationnye tekhnologii v istotricheskikh issledovanii akh i obrazovanii. St. Katharinen, pp. 5-18.

Tiele, C.P. (1897). *Elements of the science of religion*. W. Blackwood and Sons. Edinburgh.

The New English Bible. (1971). Numbers 1:1-10. New York: Oxford University Press.

The New English Bible. (1971). Luke 2:1-5. New York: Oxford University Press.

Thiel, U., Brocks, H., Dirsch-Weigand, A., Everts, A., Frommholz, I., and Stein, A. (2005). Queries in Context: Access to Digitized Historic Documents in a Collaboratory for the Humanities.

Tosh, J. (2010). *The Pursuit of History: Aims, Methods, and New Directions in the Study of History* (5th ed.). Pearson Education: Harlow.

UK Data Service. (2012). Census Support. Accessed on 08-02-2016. https://census.ukdataservice.ac.uk

Van Dalen, D. B. (1979). *Understanding educational research*. McGraw-Hill. New York.

Van de Camp, M., and Van den Bosch, A. (2011). A link to the past: Constructing historical social networks. In A. M. A. Balahur E. Boldrini and P. Martinez-Barco (Eds.), *Proceedings of the ACL-HLT Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA-2011), Association for Computational Linguistics.* pp. 61–69. Stroudsburg, PA, USA.

Van den Akker, C., Legêne, S., Van Erp, M., Aroyo, L., Segers, R., der Meij, L., Jacobs, G. (2011). Digital Hermeneutics: Agora and the Online Understanding of Cultural Heritage. In

Proceedings of the 3rd International Conference on Web Science (WebSci 2011). pp. 1–7.

Van der Bie, R. (2014). Tot opheffing van het volksleven. Mensen achter het CBS, ca. 1850–1900. *Centraal Bureau voor de Statistiek*

Van der Bie, R. (2007). 'Licht, lucht en vrijheid. De kwaliteit van de huisvesting (1899-1947)'. *Twee eeuwen Nederland geteld*, 129-151. DANS – Data Archiving and Networked Services.

Van der Meer, A. and Boonstra, O. (2006). *Repertorium van Nederlandse gemeenten 1812-2006.* Data Archiving and Networked Services.

Van Dijk, H. and S.W Verstegen. (1988). *Dienstverlening in Nederland En Duitsland: Tussen Eerste Wereldoorlog En Welvaartsstaat (1920-1960)*. Noord-Hollandsche Uitgevers Maatschappij. Amsterdam.

Van Eijl, C.J. and Lucassen, L.A.C.J. (2001). 'Vreemdelingen in de Nederlandse volks- en beroepstellingen (1899-1971)'. *Nederland een eeuw geleden geteld: een terugblik op de samenleving rond 1900*, pp. 65-88. Stichting Beheer IISG.

Van Hage, W. R., Malaisé, V., Segers, R., Hollink, L., and Schreiber, G. (2011). Design and use of the Simple Event Model (SEM). *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, *9*(2), pp.128–136. Retrieved from http://www.websemanticsjournal.org/index.php/ps/article/view/190

Van Leeuwen, M.H.D., Maas, I., and Miles, A. (2002). *HISCO: Historical International Standard Classification of Occupations*. Leuven University Press.

Van Maarseveen, J. (2002). *Algemene tellingen in de twintigste eeuw: De methode van onderzoek bij personen en bedrijven*. Centraal Bureau voor de Statistiek.

Van Maarseveen, J. (2008). Dutch Occupational Censuses 1849-1971/2001. A component of the Population Census. *Netherlands Central Bureau of Statistics.* Voorburg, The Netherlands.

Van Maarseveen, J. (2003). De virtuele volkstelling en het sociaal statistisch bestand. Een verslag van de conferentie gehouden in Amsterdam op 11 november 2003. CBS

Van Maarseveen, J. G.S.J. (2002). 'Intrekking Volkstellingenwet. Registertellingen; op weg naar volkstellingen nieuwe stijl 1979-2000', in: J.G.S.J. van Maarseveen (ed.), *Algemene tellingen in de twintigste eeuw. De methode van onderzoek bij personen en bedrijven*. pp. 89-114. Centraal Bureau voor de Statitistiek.Voorburg/Heerlen.

Van Maarseveen, J., and and P.K. Doorn (2001). *Nederland een eeuw geleden geteld. Een terugblik op de samenleving rond 1900*. Stichting Beheer IISG. Amsterdam.

Van Poppel, F.W.A. (2001). 'De groei van de Nederlandse bevolking in de afgelopen eeuw'. *Nederland een eeuw geleden geteld: een terugblik op de samenleving rond 1900*, pp. 65-88. Stichting Beheer IISG. Amsterdam.

Venetis, P, Halevy, A., Madhavan, J., Paşca, M., Shen, W., Wu, F., Miao, G. and Wu, C. (2011). Recovering Semantics of Tables on the Web. *Proceedings of the VLDB Endowment (PVLDB)*, 4(9), pp. 528–538, http://dx.doi.org/10.14778/2002938.2002939.

R. Verhoef. (1981). 'The history of population registration and demographic data collection in the Netherlands'. *National population bibliography of the Netherlands 1945–1979*, ed. by H.G. Moors. The Hague.

W3C. (2015). https://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance# A_Working_Definition_of_Provenance

Wang, S., Schlobach, S., and Klein, M. C. A. (2011). Concept drift and how to identify it. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, *9*(3), pp. 247–265.

Ward, D. J. H. (1909). *The classification of religions: different methods, their advantages and disadvantages*. The Open court publishing company. Chicago.

Wisselgren, M.J., Edvinsson, S., Berggren, M., and Larsson, M. (2014). Testing methods of record linkage on Swedish censuses. *Historical Methods*, 47(3), pp. 138-151.

Witte, R., Krestel, R., Kappler, T., and Lockemann, P. C. (2010). Converting a Historical Architecture Encyclopedia into a Semantic Knowledge Base. *IEEE Intelligent Systems*, *25*(1), pp.58–67. http://doi.ieeecomputersociety.org/10.1109/MIS.2010.17

Wittek, P., and Ravenek, W. (2011). Supporting the Exploration of a Corpus of 17th-Century Scholarly Correspondences by Topic Modeling. In B. Maegaard (Ed.), *Supporting Digital Humanities 2011: Answering the unaskable*. Copenhagen, Denmark.

Wolters, A., Woollard, M. (2014). Restoring and preserving historic data for research purposes. *IASSIST* 2014, Toronto

Yakout, M, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri. (2012). InfoGather: Entity Augmentation and Attribute Discovery By Holistic Matching with Web Tables. *Proceedings of the 2012 ACM SIGMOD Internatonal Conference on Management of Data*, pp. 97–108. New York