Historical Methods: A Journal of Quantitative and Interdisciplinary History

ISSN: 0161-5440 (Print) 1940-1906 (Online) Journal homepage: http://www.tandfonline.com/loi/vhim20

The Aggregate Dutch Historical Censuses

Ashkan Ashkpour, Albert Meroño-Peñuela & Kees Mandemakers

To cite this article: Ashkan Ashkpour, Albert Meroño-Peñuela & Kees Mandemakers (2015) The Aggregate Dutch Historical Censuses, Historical Methods: A Journal of Quantitative and Interdisciplinary History, 48:4, 230-245, DOI: 10.1080/01615440.2015.1026009

To link to this article: http://dx.doi.org/10.1080/01615440.2015.1026009

HISTORICAL METHODS, October–December 2015, Volume 48, Number 4

Routledge
Taylor & Francis Group

The Aggregate Dutch Historical Censuses

Harmonization and RDF

ASHKAN ASHKPOUR
International Institute of Social History

School of History, Culture and Communication
Erasmus University

ALBERT MEROÑO-PEÑUELA
Department of Computer Science
VU University Amsterdam

Data Archiving and Networked Services

KEES MANDEMAKERS
International Institute of Social History

School of History, Culture and Communication
Erasmus University

Abstract. Historical censuses have an enormous potential for research. In order to fully use this potential, harmonization of these censuses is essential. During the last decades, enormous efforts have been undertaken in digitizing the published aggregated outcomes of the Dutch historical censuses (1795–1971). Although the accessibility has been improved enormously, researchers must cope with hundreds of heterogeneous and disconnected Excel tables. As a result, the census is still for the most part an untapped source of information. The authors describe the main harmonization challenges of the census and how they work toward one harmonized dataset. They propose a specific approach and model in creating an interlinked census dataset in the Semantic Web using the Resource Description Framework technology.

Keywords: harmonization, historical censuses, historical demography, RDF, Semantic Web, social and economic history

Address correspondence to Ashkan Ashkpour, International Institute of Social History, Cruquiusweg 31, Amsterdam 1019 AT, Netherlands. E-mail: ashkan.ashkpour@iisg.nl

Introduction

Censuses are taken regularly by governments throughout history to gain a better understanding of populations and their different characteristics such as size, age structure, household compositions, occupations, and other sociodemographic 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 nation's *population characteristics* and *needs* at a given *time* in history, providing invaluable snapshots of the state of a nation (Higgs 1996). For the period before the twentieth century, the census is one of the only large scale historical statistical data sources on population characteristics which are not strongly distorted, providing comprehensive geographical coverage (Ruggles and Menard 1995).

The first integral enumeration in the Netherlands started in 1795 under the French influence during the Batavian Republic. It took over thirty-five years before the next general

population census was organized (1829). This was based on the Royal Decree of 1828, which stated that the census should be taken every ten years. 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. The 1971 census marks the end of the traditional census in the Netherlands, which in total covered seventeen census years for almost two centuries (Den Dulk and Van Maarseveen 1999). The end of the traditional censuses has not exempted the Dutch government in its obligation to meet European regulations and to collect this type of information about its population. Currently, the census is harvested digitally from the municipal registrars.[1]

Unfortunately, because of the existence of the population registers from 1850 onward, 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. The number of volumes depends on the specific census year. 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. Physical presence and cumbersome manual efforts were required in order to extract meaningful data from the census. In order to provide better access to and use of the census data, major efforts have been taken in the digitization of the census, starting in 1996. From this year onward, the Dutch Statistics (CBS) and the institute Data Archiving and Networked Services (DANS) worked together in digitizing the books with the aggregative results of the censuses from 1795–1947 to improve the accessibility of the 1960 and 1971 micro datasets. The first step in the digitization process was to scan the books and publish them as images in order to provide better access to the historical census data and also to preserve this material for the future. Information which previously was poorly accessible to researchers (e.g., via different university libraries, Dutch Statistics, and different institutions) was now made available by way of Internet and CD-ROMs. However, the images are very difficult to handle. A single table in the census can be represented by hundreds of images. The next step, therefore, was the shift from *medium* to *content* conversion, in which the images were manually transcribed into Excel workbooks. During the transcription process, the choice was made to represent the census tables in a source-oriented manner, meaning that the tables in Excel should resemble the tables in the books as closely as possible (including the presentation of the tables). As a result, the researchers ended up with 2,249 Excel tables instead of an integrated harmonized dataset. These Excel tables are the basis for the next steps in the digitization process of the Dutch historical censuses and form the starting point for our harmonization efforts.

Although now digital and computer processable, the aggregate historical Dutch census is still not being used to its full potential. Besides the data representation limitations of having thousands of heterogeneous and unconnected Excel tables, another common problem relates to data harmonization. The disconnected Excel tables present many different classification systems which must be *standardized* to allow temporal comparisons. Next to structural problems, we find all kind of inconsistencies not only in the structure of our tables, but also in the data itself as both source and digitization errors have been introduced at different stages.

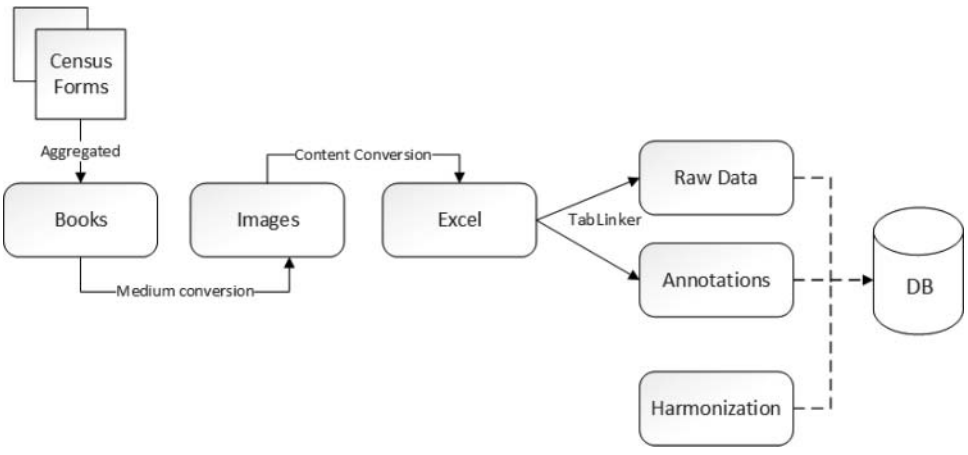


FIGURE 1. Census digitization process.

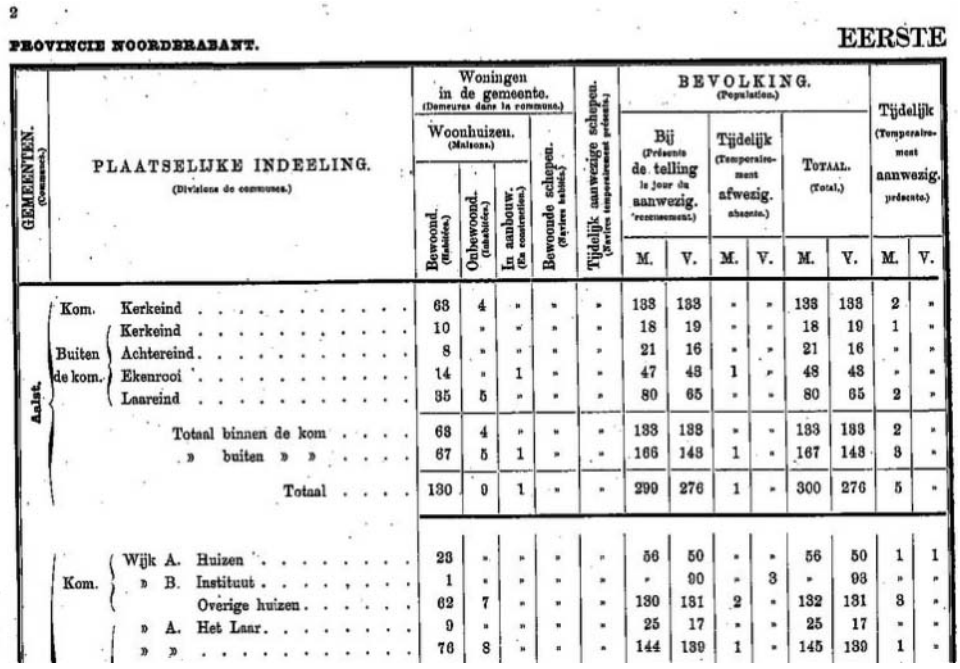
Figure 1 shows the various stages of the digitization process from the census forms to Excel tables, into a harmonized database. The last step will be done by using Semantic Web technologies. In order to move toward a database system of the Netherlands' aggregate census statistics which can be queried uniformly, we apply a specific knowledge representation model from the Semantic Web: the Resource Description Framework (RDF). The RDF framework allows us to provide better access to and use of the historical censuses using linked data principles.

This article consists of three main parts. In the first part, we will elaborate on the historical background and varying structures of the Dutch census dataset, and the way it has been handled in the digitization process so far. In the second part, we will delve into specific problems of harmonization. The third part presents our specific approach, proposing a solution for going from the Excel tables to an interlinked harmonized dataset. We will describe *how we converted* the Excel tables into RDF using a specific method which is suitable for heterogeneous tables with different hierarchies; *how we modelled* this using a three-tier architecture, separating the raw data from the annotations and harmonization layers; and *how we harmonized* and *queried* our dataset after this conversion to allow longitudinal comparisons.

Dutch Census Dataset

When referring to the published results of the Dutch historical census data over the years, we must distinguish three main types of aggregate census data: population, occupation, and housing data. Published tabulations on the population span the entire range of our historical census dataset, whereas the occupational census tables were only published for the censuses of 1849, 1859, and 1889 onward. Until 1930, information on the housing statistics of the Netherlands was published as part of the population census which contained some tables with "housing statistics." The first official housing census was introduced in 1947 and was linked to the population census. In 1956, the housing shortage and need for data about the housing market called for a new housing census, and it was conducted separately, independent of the population census. The last official housing census was held in 1971, again together with the population census.

As a first effort in both preserving and providing better access to the original census books, different digitization projects were undertaken by Dutch Statistics and DANS (Doorn, Jonker, and Vreugdenhil 2001; Van Maarseveen 2008). These projects concentrated on the digitization of the census books and resulted in around 22,000 images, representing all tables with the census results. Although more accessible and better 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, which as such are also quite unreadable on a normal screen without having them enlarged four or five times. The second step in the digitization of the census focused on *content conversion*. Accordingly, the images were converted into computer processable files, that is, *Excel files*. Experiments with Optical Character Recognition (OCR) did not lead to

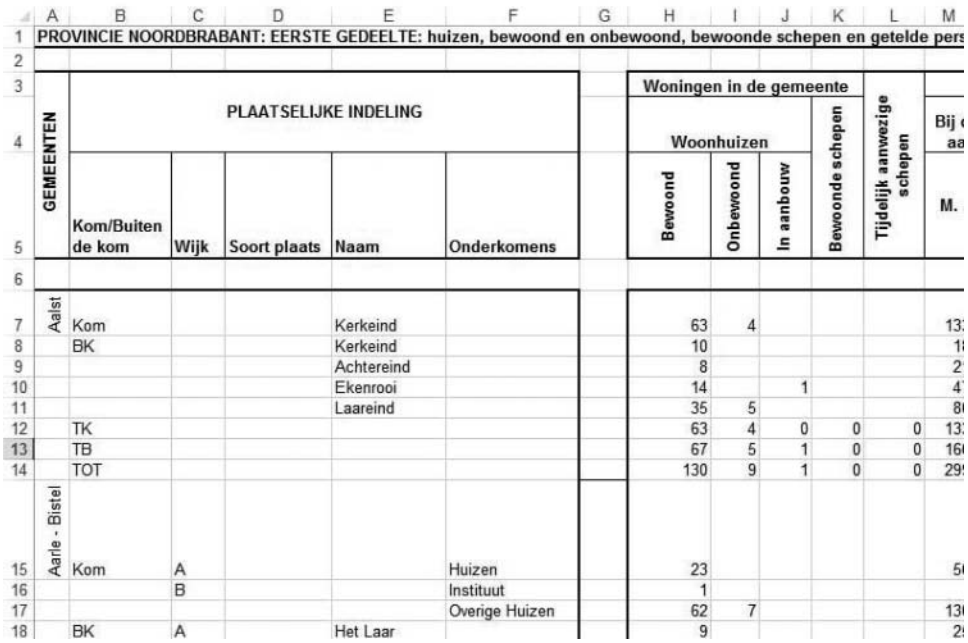


2

PROVINCIE NOORDBRABANT. EERSTE

GEMEENTEN. (Communes.)	PLAATSELIJKE INDEELING. (Divisions de 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 in de nacht van den 31sten December op den 1sten Januari aanwezig. (Présents au 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	»	»	133	133	2	»
	Buiten de kom. Kerkeind	10	»	»	»	»	18	19	»	»	18	19	1	»
	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	143	3	»
	Totaal	130	0	1	»	»	299	276	1	»	300	276	5	»
	Kom. Wijk A. Huizen	23	»	»	»	»	56	50	»	»	56	50	1	1
	» B. Instituut	1	»	»	»	»	»	90	»	3	»	93	»	»
	Overige huizen	62	7	»	»	»	130	131	2	»	132	131	3	»
	» A. Het Laar	9	»	»	»	»	25	17	»	»	25	17	»	»
	» »	76	8	»	»	»	144	139	1	»	145	139	1	»

FIGURE 2. Example of a scanned table image.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	PROVINCIE NOORDBRABANT: EERSTE GEDEELTE: huizen, bewoond en onbewoond, bewoonde schepen en getelde pers												
2													
3		PLAATSELIJKE INDELING						Woningen in de gemeente					Bij
4	GEMEENTEN							Woonhuizen			Bewoonde schepen	Tijdelijk aanwezige schepen	aa
5		Kom/Buiten de kom	Wijk	Soort plaats	Naam	Onderkomens		Bewoond	Onbewoond	In aanbouw			M.
6													
7	Aalst	Kom			Kerkeind			63	4				13
8		BK			Kerkeind			10					1
9					Achtereind			8					2
10					Ekenrooi			14		1			4
11					Laareind			35	5				8
12		TK						63	4	0	0	0	13
13		TB						67	5	1	0	0	16
14		TOT						130	9	1	0	0	29
15	Aarle - Bistel	Kom	A			Huizen		23					5
16			B			Instituut		1					
17						Overige Huizen		62	7				13
18		BK	A		Het Laar			9					2

FIGURE 3. Example of a table converted to Excel from images.

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 et al. 2001). Figures 2 and 3 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 from the website http://www.volkstellingen.nl.

The main principle in the conversion process from images to spreadsheets was to represent the source as closely as possible. While this approach is typically a golden rule in constructing microdata, in the case of reproducing aggregate statistics, it can be a problem. This source-orientated process means that no efforts were undertaken to harmonize the data and structure of the census tables. Each Excel table applies to a certain year, specific region (municipality, province, and national total), and specific census type (i.e., population, occupation, or housing census). In total, the electronic historical Dutch census consists of 2,249 Excel tables with aggregated data waiting to be aligned and harmonized in order to allow studies over time and space.

Table 1 shows the distribution of the number of *tables with aggregated data* and *annotations* per census year (1795–1956). These annotations have different types of meaning and were made in different ways. They 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 approach, the original figures in the tables were generally not changed (although we found examples that indeed the source was corrected). We sometimes find annotations as comments in a cell, in another sheet, or even as replaced values. Most of the annotations in the census are textual (whether a comment or interpretation), and only a small number are actual corrections to the data (numerical). All in all, we deal with 33,283 annotations in the Excel files of which it is not

Table 1. Census Digitization Statistics

Year	Tables	Annotations
1795	28	100
1830	17	71
1840	60	27
1849	94	75
1859	183	4,896
1869	226	321
1879	985	516
1889	166	14,349
1899	76	2,594
1909	138	3,381
1919	4	224
1920	48	5,396
1930	32	1,112
1947	133	83
1956	59	138
Totals	2,249	33,283

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 have created a bottom-up standard classification system from the current annotations in the census and propose a specific model in RDF in order to organize the annotations and deal with future changes in a consistent manner (see example in RDF section).

Harmonization: Problems and Solutions

In general, the structure of a census is subject to change from year to year due to different systems or 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 activity. Censuses which are collected and digitized over long periods of time have significant limitations and are hampered with evolving variables, structures, observation methods, questions, processing methods, and classification systems which make it difficult to fully reap the potential of the census (Van Maarseveen 2008; Ruggles and Mennard 1995; Putte and Miles 2005). The Dutch historical census is no different and shares many of these problems, and even worse, it provides only *aggregated data* in tabular form to work with.

One of the first steps in the harmonization of the Dutch census is to eliminate unnecessary complexity by converting the content of the 2,249 Excel tables into a unified dataset in the form of a database system which can be queried. In a very straightforward approach, we have departed from the Excel tables and converted our dataset to RDF and stored it in a RDF database system (called a Triplestore) which we aim to build on. More information on how we did this will be elaborated in the next section (see Census to RDF section).

In order to move toward a harmonized aggregate census database, we must 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, in the following sections we will describe the key challenges we face in the harmonization of the Dutch census: working with aggregated data, changing variables, creating variables, structural heterogeneity, inconsistencies, and changing classifications.

Aggregated Data

Statistical census data are typically presented on aggregated levels. This aggregation answers the information needs of the public, politicians, government, and so forth 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., Sweden, 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 microdata, harmonizing aggregated census data on a diachronic basis is hampered as it is not possible to simply build or rebuild a classification. Unlike many similar census harmonization efforts (see section: Comparable Studies), we cannot reconstruct the (classification) systems at a microlevel to suit our needs. Our harmonization work therefore concentrates on two problems: First of all, we must harmonize the variables and values over time, and secondly we must harmonize the totals from the several hierarchical layers in which the census results are published. The second problem arises when the national total of some specific variable is not the same as the sum of the provincial totals for that variable. Similarly, we sometimes find that the sum of the number of inhabitants in all municipalities for a certain province does not match with the total number of inhabitants given for that province. The lack of microdata necessitates the use of a *combination* of statistical approaches with regardsto harmonization of aggregated data. Considering this, we are constrained to provide higher level aggregations, create new variables, and 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. Documentation is provided both at cell and variable level in order to allow the users to judge the appropriateness of the transformations for their research.

Changing Variables

Classifications systems are used in the census in order to categorize the various *variables* and 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. Radical changes in the classification systems and coding 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 *level of detail* provided. We find some variables which stay more or less the same over time, such as gender, marital status, housing types, and so on, 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*) to reflect 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 this often with geographical variables, such as municipalities which have changed significantly over the course of time in the Netherlands (Van der Meer and Boonstra 2006). For example, the composition of the municipality of Rotterdam changed significantly between the late nineteenth and mid-twentieth century, having nine major changes between 1886 and 1941.

Examples of changing variables across time can also be found throughout the "occupational census" due to innovation (i.e., specialization, differentiation, etc.). During its lifespan, 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 historical censuses: 1849–59, 1889–1909, and 1920–47.

Another example of variables and classification systems which evolved significantly over time is religious denomination. 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" ("Quakers").

Creating Variables

The meanings of variables and concepts in our dataset are subject to change from census to census. While in some cases it is simply different labeling of a variable, we also find distinctions in variables which are much more difficult to harmonize. When working with aggregated data, the creation of new variables is a common solution used in the harmonization process. For example, in the case of the housing type classification, we have very specific detail on how many people were counted in barracks (e.g., "Kazerne der Maréchaussée," "Artilleriekazernes," and so on) or forts (e.g., "Fort Isabelle" and "Fort Kijkduin"). As we do not have this detailed information for all years, we combine these housing types according to the function they performed and create a new higher level variable "Military Buildings." However, problems such as changing *age categories* require different statistical methods based on estimations. As we deal with aggregated data rather than microdata, we cannot simply reconstruct new "age ranges" to allow comparisons across time. New variables must be constructed to make age categories which cover all the census years. Whereas in some cases simple addition could be sufficient, this is not always the case. For example, in order to make the age ranges "14–18" and "19–20" comparable with "14–15" and "16–20," a typical solution could be by (a) making a new category "14–20" and (b) by interpolating three new categories: "14–15," "16–17," and "18–20." Although these "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 provide the user with different solutions for the same variable. In the case of the occupational census, DANS had already connected our dataset to external classification systems such as HISCO (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 fine-grained detail of the census. In this context, we must create standard classification systems for housing types, religious denominations, and lower level classifications of occupations and of municipalities such as neighborhoods, areas, and so on.

Although we provide 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 must perform 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*

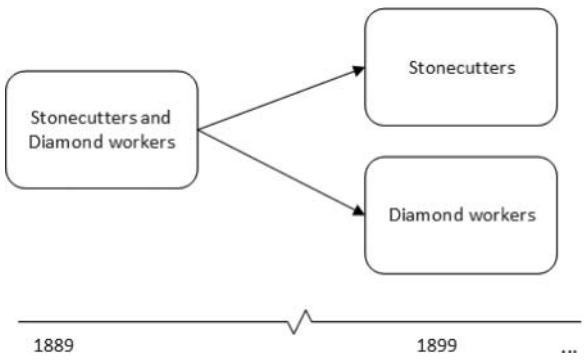


FIGURE 4. Splitting of an occupational class.

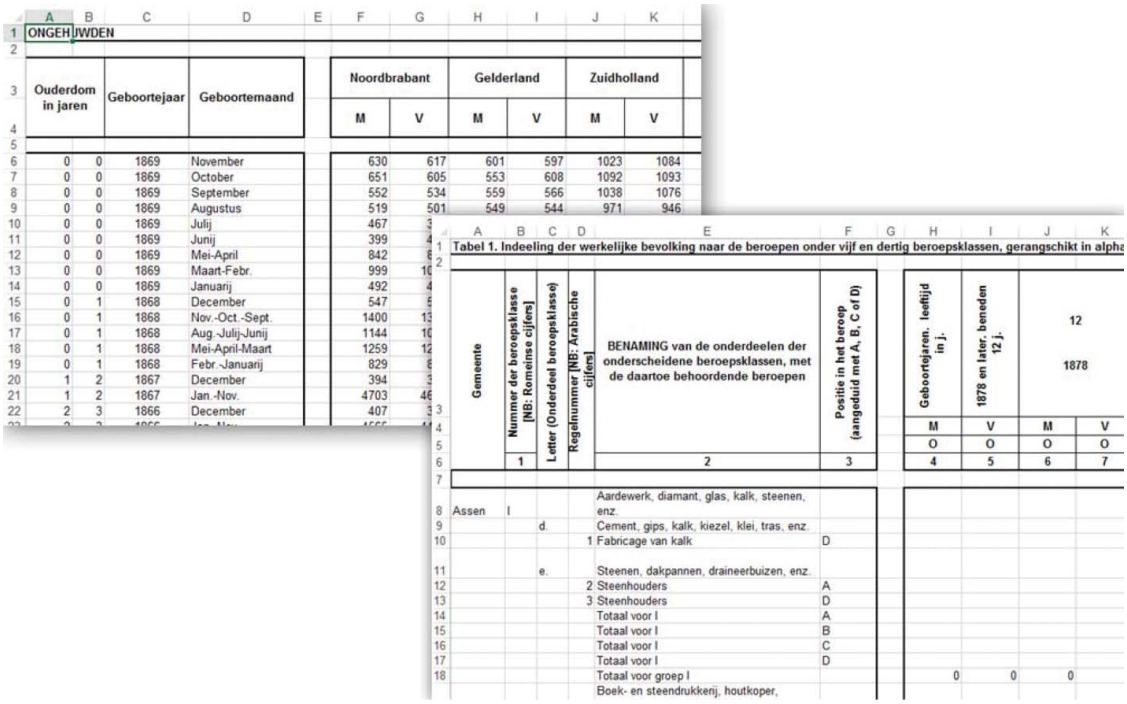
up or *dividing* a category according to a certain ratio could suffice. This is, for example, the case of the diamond workers in the occupational census of 1899, in which they received their own category and were no longer grouped with stonecutters as in 1889 (see Figure 4). Accordingly, in this specific 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 in the form of flags (will be elaborated in the following sections) 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.

Structural Heterogeneity

During the digitization of the historical censuses, the choice was made to apply a 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 toward one system, we must 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 5 for an example of two Excel tables with distinct 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



ONGEHUWDEN

Ouderdom in jaren	Geboortejaar	Geboortemaand	Noordbrabant		Gelderland		Zuidholland	
			M	V	M	V	M	V
0	0	1869	November	630	617	601	597	1023	1084
0	0	1869	October	651	605	553	608	1092	1093
0	0	1869	September	552	534	559	566	1038	1076
0	0	1869	Augustus	519	501	549	544	971	946
0	0	1869	Julij	467	3				
0	0	1869	Junij	399	4				
0	0	1869	Mei-April	842	8				
0	0	1869	Maart-Febr.	999	10				
0	0	1869	Januarij	492	4				
0	1	1868	December	547	5				
0	1	1868	Nov.-Oct.-Sept.	1400	13				
0	1	1868	Aug.-Julij-Junij	1144	10				
0	1	1868	Mei-April-Maart	1259	12				
0	1	1868	Febr.-Januarij	829	8				
1	2	1867	December	394	3				
1	2	1867	Jan.-Nov.	4703	46				
2	3	1866	December	407	3				

Tabel 1. Indeeling der werkelijke bevolking naar de beroepen onder vijf en dertig beroepsklassen, gerangschikt in alpha

Gemeente	Nummer der beroepsklasse [NB: Romeinse cijfers]	Letter (Onderdeel beroepsklasse)	Regelnummer [NB: Arabische cijfers]	BENAMING van de onderdeelen der onderscheidene beroepsklassen, met de daartoe behoorende beroepen	Positie in het beroep (aangeduid met A, B, C of D)	Geboortejaren. Leeftijd in j.		12 1878	
						M	V	M	V
						O	O	O	O
	1			2	3	4	5	6	7
Assen	I			Aardewerk, diamant, glas, kalk, steenen, enz.					
		d.		Cement, gips, kalk, kiezel, klei, tras, enz.					
			1	Fabricage van kalk	D				
		e.		Steenen, dakpannen, draineerbuizen, enz.					
			2	Steenhouders	A				
			3	Steenhouders	D				
				Totaal voor I	A				
				Totaal voor I	B				
				Totaal voor I	C				
				Totaal voor I	D				
				Totaal voor groep I		0	0	0	
				Boek- en steendrukkerij, houtkoper,					

FIGURE 5. Example of different table structures.

dataset, without compromising valuable data. Trying to force an overall data model on the 2,249 Excel tables would practically mean that we must harmonize everything to the broadest category, resulting in the loss of valuable subcategories 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. 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.

By choosing for a more integral and harmonized approach, we work in the lines of Esteve and Sobek (2003) and define census harmonization as the creation of a unified, consistent data series from dissimilar census data.

Dealing With Inconsistencies

It will be clear that besides changing variables and classification systems, the structural heterogeneity of the tables and the aggregated character of the data in itself may cause major inconsistencies when making one system 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 a source annotation has 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 these inconsistencies, we must clean, correct, enrich, standardize, and even restructure the data to have an *acceptable* dataset to do research with. All these "improvements" to the data are part of the harmonization package and are sometimes even necessary before being able to continue in the harmonization process itself. 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, etc.). As no consistent logic is applied, it is very difficult to extract the right data without extensive manual input.

We use different scripts to manage these inconsistencies.[2] 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. 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. These different methods are applied to check the quality of the data and improve the inconsistencies. To further test our methods, we harmonized a subset of the population census from 1859–89 in the form of mini projects which we designed specifically to explore the possibilities of harmonizing aggregate historical census data in RDF.

Harmonization and the RDF Approach

In this section, we elaborate on our approach using RDF to model and harmonize the aggregate historical censuses over time. We explain our motivation for using RDF as the modeling technique, how we harmonize the aggregate censuses over time in RDF, and the three-tier model we created to provide a flexible harmonized census database.

The Semantic Web and RDF

Envisioned in 2001, the Semantic Web (Berners-Lee, Hendler, and Lassila 2001) was conceived as an evolution of the original World Wide Web, which is built essentially on documents. Most of the contents of the Web are designed for humans to read, but not for computer programs to process meaningfully. Computer programs are able to parse the source code of Web pages to extract layout information and text, but there are no mechanisms to process their semantics (Meroño-Peñuela et al. 2015). The "Semantic" Web enables the sharing of content from databases and other structured data sources which are not directly published on the Web. The Semantic Web "is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee et al. 2001, 37).

More practically, the Semantic Web is also 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 RDF[3] is the basic layer on which the Semantic Web is built. W3C 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."[4] It is used as a conceptual description method in computing. Entities of the world are represented with *subjects* and *objects* while the relationship between the two is represented with *predicates* connecting them (e.g., "Amsterdam" "is located in" "The Netherlands"). Hence, RDF is a knowledge representation system where facts and their properties are expressed as subject-predicate-object sentences known

```
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. Example of SPARQL enriching the data from other RDF sources.

as *triples* (e.g., Amsterdam-isLocatedIn-TheNetherlands), and all connected have the form of a graph. Finally, all unique subjects, predicates, and objects are assigned a Uniform Resource Identifier (URI) that uniquely identifies them on the Web. Once converted and published, RDF data can be queried online through the query language SPARQL[5] (SPARQL Protocol and RDF Query Language).

Using RDF as a knowledge representation model, we created a *one to one* copy of the structure and contents of the Excel files in the form of a (graph) database and separate the harmonization process from the data itself (Mandemakers and Dillon 2004; Meroño-Peñuela et al. 2012). We facilitate the different harmonization views/interpretations by creating a three-tier architecture in RDF where we separate the raw *data* from the *harmonization* and *annotations* in the census. 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 of the Dutch historical census.

There are several reasons why RDF is chosen as the data system in which we model, publish, and query the Dutch census dataset. 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 disconnected 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 table must conform to the same schema; therefore, these tables can be represented with diverse RDF graphs that match their diverse structure, without constraints on meeting an overall agreed schema. This is especially useful to extend and particularize descriptions of resources; for instance, variables can be more concretely defined with the specificities and constraints 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 (its query language) were designed to merge disparate sources of data on the Web. For example, the following SPARQL query illustrates how the linkage between the Dutch historical censuses in RDF and other sources of linked data on the Web is used to extend information on Dutch municipalities, comparing their 1889 and current populations from the CEDAR and DBPedia's[6] linked data endpoints (see Figure 6), respectively.

The Dutch census case enables us to 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. With such a linked dataset, extended and enriched census information can be retrieved combining data from the linked sources (e.g., number of workers per occupation and year versus number of labour strikes per occupation, year, and municipality or region).

Comparable Studies

Over the past years, different efforts have been undertaken using RDF technology for greater census utilization. The 2000 U.S Census (Tauberer 2007) was converted to RDF providing population statistics on various geographic levels.[7] Although not historical and harmonized only for that specific census, it deals with the same challenges in finding an appropriate data model to represent the census data in RDF. In Canada, the Canadian health census uses linked open data

(LOD) based on RDF in order to provide greater access to and usage of the data, and to promote greater interoperability which cannot be achieved with conventional data formats (Bukhari and Baker 2013). Using a scalable and interoperable format such as RDF is intended to make their data reusable across different platforms. In another comparable approach, in the context of a national large-scale project to manage sociodemographic data in Greece, Petrou and Papastefanatos (2014) applied LOD technologies to the Greek population census of 2011. Their goal is similar to that of the Canadian health census, to publish "traditional" datasets in RDF and thus allow easier access and use of the census by third parties. The 2001 Spanish Census project is another advocate of applying LOD technologies such as RDF to the census and encouraging the development of the open government philosophy (Fernández et al. 2011). Using microdata from the 2001 population census, Fernández and colleagues (2011) proposed a solution for converting the data into open formats allowing greater discoverability, accessibility, and integration of the data, which is a recurrent theme in all of the mentioned projects.

All these projects have merely harmonized data within the domain of each census year, have used microdata as a starting point, and have focused mainly on publishing the data. In the following, we will explain how we have used RDF in a novel way and propose a *three-tier model* to harmonize the data over time.

A Three-Tier Model

In order to deal with the challenges of the census, we model our dataset in a three-tier architecture according to the multitier architecture principles where layers are logically separated. In our model, the architecture consists of the *harmonization* layer, the *raw data* layer, and the *annotations* layer (see Figure 7). The dependencies between the layers are represented in the figure with directional arrows. An arrow from *A* to *B* means that structure and data from *A* must be linked to structure and data from *B*.

We separate the data in this way 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 trace 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 in the future. Accordingly, we have designed a workflow in order to feed new annotations into our three-layered model. We also allow suggestions for changes to the data via an online interface (to control the quality of the data, the workflow 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 deal with 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 it may need to be customized according to multiple research requirements. Storing the different harmonization practices in a separate layer allows us to modify the harmonization procedures as we go, without affecting the underlying raw data. Moreover, due to the ambiguity of some harmonization practices, this approach allows us to provide several solutions to each particular problem.

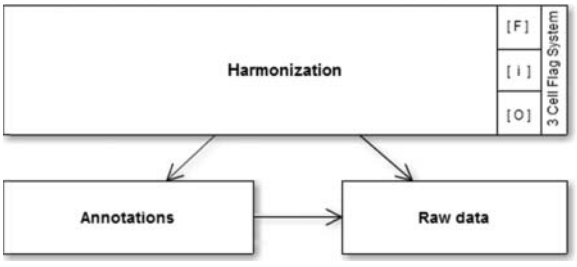


FIGURE 7. Harmonization three-tier model.

Raw Data Layer

The *raw data* layer consists of a *one to one* copy of the original Excel sheets (see Figure 3). The 2,249 Excel tables with their different structures are stored in this layer in the form of RDF graphs. Since the data contained in a census table are 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.[8] RDF Data Cube is the *standard* for modeling and publishing multidimensional data, such as statistics, in the Semantic Web.

While the *layout* and *structure* of the Excel tables differ significantly across our dataset, they contain the same basic structure of three areas: 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 must specify for each table where the columns, rows, and content area *start* and *end*. This is done by way of so-called bounding boxes by which we 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. Exploiting this common characteristic of the tables allows us to apply the same approach in converting all Excel tables to RDF.

Harmonization Layer

Harmonizing the aggregate Dutch historical censuses draws upon a combination of different harmonization practices including resolving inconsistencies, and data cleaning,

correcting, and restructuring, 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 creating standard vocabularies, constructing variables across the different censuses, creating new variables and values, and connecting them to existing classification systems such as HISCO (for occupational variables) and the Amsterdam Code (classification for municipalities in the Netherlands). For this connection, we must convert these classifications into RDF. Although we use standard vocabularies whenever possible, the censuses often require that we create our own classifications and vocabularies. We store all these types of created data (mutations, new variables, classifications, etc.) in the harmonization RDF layer, which can be enriched and modified as a continuous iterative process without compromising the underlying data.

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.

When we combine two variables to create a higher level aggregation for harmonization purposes, we in fact create a new level of the *Three Cell Flag System*. Building on the example illustrated in Figure 4, where we harmonized the "stonecutters" and "diamond workers" of 1899 into one group to make it comparable with the census of 1889, we combined 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. Also, 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.

Harmonizing in Four Stages

Practically, we harmonize the dataset in four stages. Firstly, we define which standard variables and values we will use for the raw variables and values which we find in the censuses (in RDF Data Cube terminology, respectively, *dimensions* and *codes*). We try to use as much as possible existing standardized variables, for example, the already mentioned Amsterdam code for Dutch municipalities (Van der Meer and Boonstra 2006). Secondly, we map the original variables and values of the census tables with the standard ones. Thirdly, we define transformations to create new variables or values, for example, the creation of new age categories or the splitting of occupational (sub) classes (see Figure 4). Fourthly, we use the outcomes of all previous stages to *smooth* the data into one system, harmonizing the totals of all the geographical layers with each other to achieve a consistent outcome, ensuring that the sum of all provinces for a particular variable is equal to the national totals for all variables. Finally, bringing together the richness of our three-tier model, we provide one big flat table (with all variables and census years) for expert users as well as different harmonizations of the data which can be queried and linked to other datasets online.

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.[8] The created annotation layer is linked with the raw data layer (see Figure 7). For provenance purposes, we also attach an author 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 2 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.

Table 2. Annotation Classification Dutch Census Based on a Subset of the Data

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

Census to RDF

Currently, we have progressed to the point of converting all raw datasets from the census Excel spreadsheets into an RDF triple dataset. The first stage in moving toward a harmonized database from the Excel sheets begins with the conversion of the tables into RDF, using a script called TabLinker[9] (Meroño-Peñuela et al. 2012). This script converts the structures of the Excel tables into an RDF graph for each census table. To maintain the structures of the tables, TabLinker must define different styles to link all cell values to the corresponding columns and rows. Using standard functionalities in Excel, we *color/style* the boxes of our data manually defining the columns, rows, and cell values of each table (see Figure 8).

This coloring process is very straightforward and creates a faithful (one to one) representation of the tables in RDF. *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 8).
- ***Data*** marks the data cells with the actual census numbers (the white colored section in Figure 8). Since all measurements in the dataset are counts, these numbers are qualified as integers (xsd:integer) during the conversion; additional metadata is also attached 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 zeroes.
- ***ColHeader*** marks the column headers of the table just above the content of the cells (the light blue colored section in Figure 8). These headers contain the values for different variables such as age ranges, marital status, sex, etc.

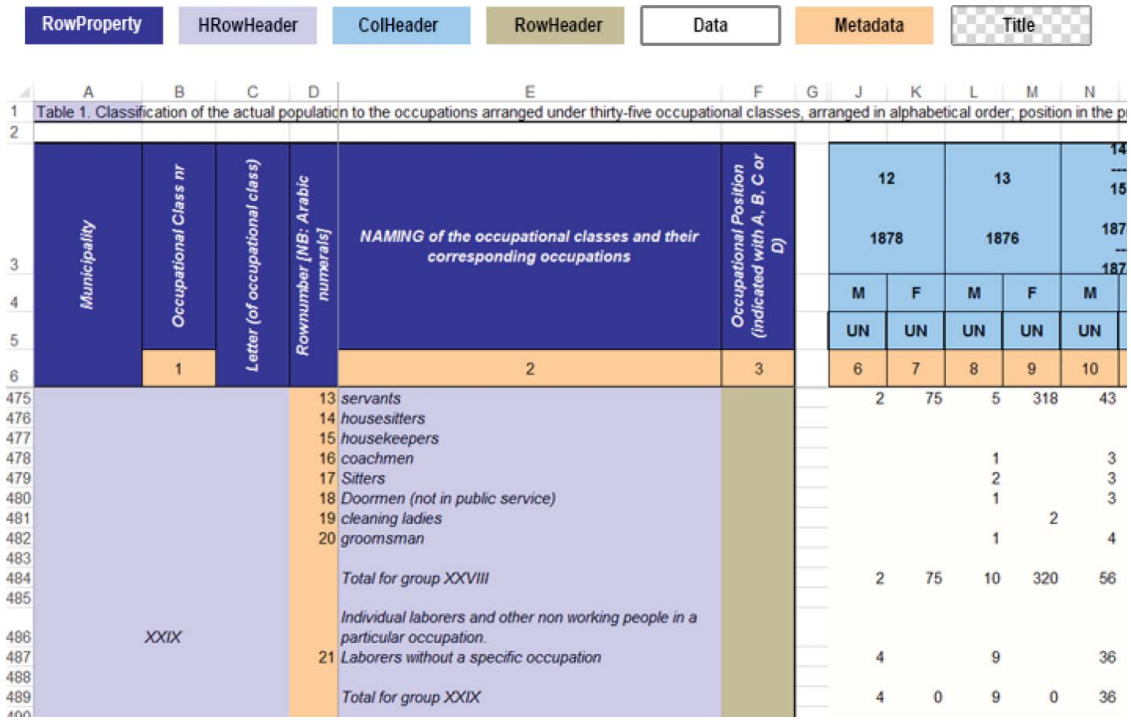


FIGURE 8. Marked census table with TabLinker (translated for illustration purposes).

- ***RowHeader*** marks cells with row headers, usually placed at the left of the table (the caramel colored section in Figure 8). These cells usually contain values for geographical variables like municipality.
- ***HRowHeader*** marks cells with hierarchical row headers (the purple colored section in Figure 8). 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 8). 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 8).

For each census table, we generate three RDF large interconnected graph-systems, shown as three layers in Figure 7. Samples of such separate RDF graphs are shown in Figures 9 and 10. Figure 9 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, and so on. In this case, the cell contains the value 1 (*"1"*^^*xsd:int*), and the headers define the content of the cell as persons of 14 or 15 years old (*:14–15_1875–1874*), being a man (*:M*), being 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*). As this is only a description for *one* single data cell in the table, having thousands of cells per table generates a much larger interconnected graph.

Similarly, the example shown in Figure 10 specifies an annotation (labeled *:y*) pointing to a cell with coordinates *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 21, 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 2), which indicates that the annotated cell contains an incorrect value (*'1'*) that should read *10.*

All these graphs are stored in an RDF database called *Triplestore*. From this Triplestore database, we will build the harmonized database to be distributed to researchers, and we will provide access for live online querying via a SPARQL endpoint.[10] 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.

Using the RDF Data

To date, the historical censuses were merely available in the form of Excel files. Presenting the generated RDF data on the Web via a SPARQL endpoint enables users and Web applications to retrieve, analyze, and visualize the historical census data, which is now available in one system. We have written client applications that query the endpoint and draw maps displaying the population of the Netherlands according to some user constraints (Meroño-Peñuela et al. 2012). We have developed query templates that allow us to systematically access the dataset in a homogeneous way. All Dutch historical census data (in our raw data layer) is now available in our SPARQL endpoint at one single Web address,[11] showing the following:

- 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.

We have already seen some examples of how these layers interact (e.g., how an annotation influences the result retrieved from the query over a data cell). More importantly, at the moment of the query, the three layers come together to give the census data all its expressive power. Client applications (i.e., applications that use such endpoint as a data source) can be developed independently by different types of users. The SPARQL endpoint can be seen, in fact, as an online database plug that any application can *leverage* via the Web. This gives researchers, historians, and developers the opportunity to build their own applications 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.

We are currently capable of retrieving any piece of information of the Dutch censuses from the raw data layer (see Figure 7). Accessing the raw data mainly allows us to purse *debugging* (detecting problems with the data) and *harmonization* as ongoing work. Practically, querying the raw data enables us to extract the needed variables, assess

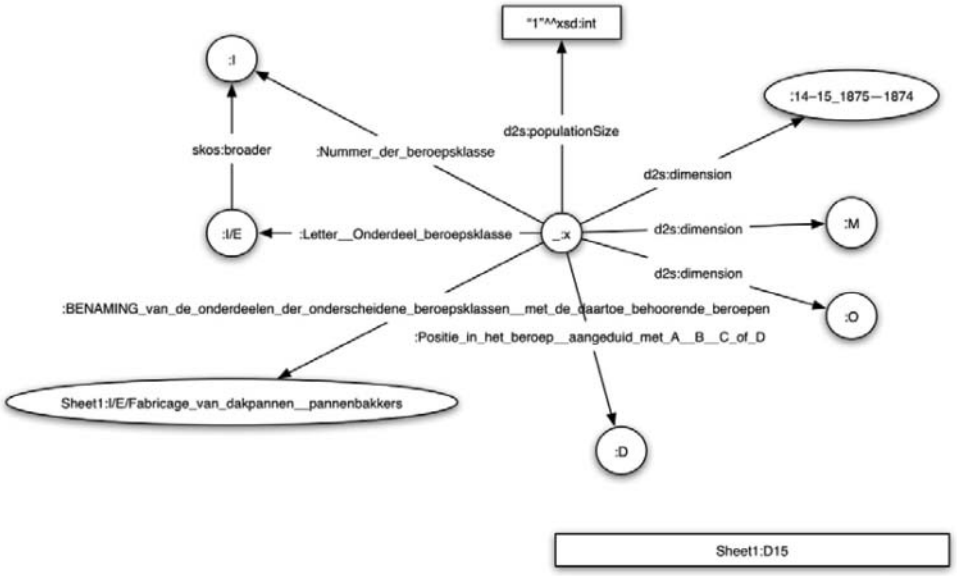


FIGURE 9. Raw data layer graph.

the quality of the data, identify already common variables across the years for classification purposes, visualize them, and detect outliers. For example, visualizing the population of Noord Brabant on the map revealed several cities, clearly falling outside of this province.[12] We can also determine where a variable breaks down over time for harmonization purposes. 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

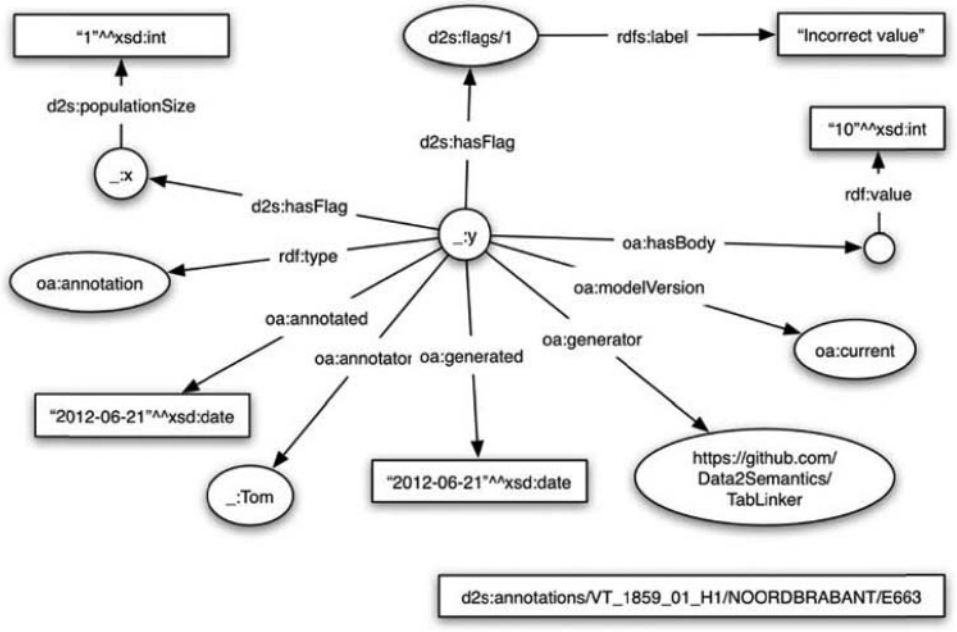


FIGURE 10. Annotation graph.

tree, that is, we can identify variable creation and extinction which thus maps common variables across the years.

We acknowledge the user perspective as an important aspect. Accordingly, we have identified three different types of users with diverse needs. Firstly, we will make the data accessible to a broad range of users such as the general public or users which are not familiar with the census data. We provide time-series and variable selection via an online interface, allowing the users to select a certain time range and some predefined harmonized variables to query, explore, and visualize the census data. Secondly, we aim to serve more advanced users such as historians, historical demographers, and sociologist, who have been working with the census for quite some time and are used to integrating the data into their own workflows and tools. We accommodate the needs of these users by providing a dump of the dataset in formats which they are familiar with for download. Finally, we also want to allow users familiar with SPARQL and RDF to query our dataset independently (both the raw data as well as the harmonized data), to be able to build their own queries, datasets, applications, and links with other datasets to build on top of our data. We provide a SPARQL endpoint via an online interface,[13] example queries,[14] the underlying data model,[15] a browsable Web interface for the RDF graphs,[16] regular RDF dumps,[17] and the complete set of developed tools and scripts[18] to implement our workflow as open source code.

Conclusions and Further Work

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 fully reaping the potential of census data in scientific research. In the Netherlands, census data have been collected for almost two centuries. Over the past twenty years, various digitization efforts have been conducted which have left us with over 2,000 heterogeneous Excel tables with aggregated census data, each with its own specific data structure. These digitized tables are our point of departure in the harmonization process. The aggregated nature of our data leads us to an approach which is different compared to other census harmonization efforts such as IPUMS. We described the challenges associated with harmonization of aggregated historical census data, identified different harmonization types (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, and so on.

In order to achieve integration, we first must transfer the Dutch Historical census data into RDF and model it according to current standards of the Semantic Web. We have developed a specific tool which deals with heterogeneous excel files (containing aggregated data), allowing us to convert the census tables into RDF in a very straightforward process without losing any information from the original tables. With this approach, we preserve valuable fine grained information contained in specific census years for researchers such as historians and historical demographers interested in the original categories, by not simply aggregating the data to higher level categories in an early stage. Moreover, as more images will be transcribed into Excel tables in the future, this tool can easily be reused to expand the current dataset. We have designed a specific model in RDF where we separate the raw data from the annotations and harmonization. We have developed standard templates and interfaces for querying the data in a uniform manner and experimented with different visualizations to explore our dataset. Through an online interface which directly plugs into our raw data layer, we have already begun to standardize variables and values, and connect them across the years. Moreover, we were able to link the municipalities and occupations in the historical censuses with existing classifications like HISCO and the Amsterdam Code for Dutch municipalities. We have also developed new standards for other variables, such housing types, annotations, and religious denominations.

We applied a specific method in RDF to model our data with all their complexity. The changing structure of the census and ambiguity of the variables requires a design which is flexible enough to allow different harmonizations of the data. To allow for this, we apply a Three Cell Flag system which takes into account the original value of the data, the interpretations we assign to the variables, whether harmonized or not, and the specification of the actions which have been undertaken to harmonize or correct the original data.

We are now on the verge of providing much better access to the historical Dutch censuses. Currently, the digitized historical censuses are published in the Semantic Web, and with some SPARQL knowledge, users are already able to query the entire census data contained in our raw data layer. The enriching of the harmonization and annotation layers is an ongoing process. Even though all the raw data are now displayed via RDF and available online, this raw database still carries the same challenges when using it. To fully reap the potential of this dataset and allow comparisons over time, it is crucial to keep building on the harmonization layer. By applying a combination of bottom-up and top-down approaches, we have already enriched the harmonization layer with spelling variants, standard variables and values, external classifications system such as HISCO and the Amsterdam Code, bottom-up historical housing, and religious classifications.

In contrast with current census harmonization efforts dealing with aggregated data (whether in RDF or with traditional approaches), we provide a model and specific approach for harmonizing historical censuses *across time*. By already presenting our raw data online, we allow third-

party users to build their own datasets, harmonization, and/or tools on top of the data. Future work will consist of further enriching the harmonization layer of our model by creating and adding our own (bottom-up) harmonization vocabularies next to connecting to spin-offs of existing systems (such as HISCO). We aim to provide a Web-based user interface which allows the users to query the data based on variable selection and time series, and the option to export these data in different formats. Although we provide all the benefits of RDF, we also consciously want to shield users from the RDF output and provide clean, *known formats* such as CSV, Relational Databases, or even round-tripping to Excel tables so users can integrate these data into their own workflows.

The transformation of the Dutch historical censuses into RDF builds on earlier digitization efforts and on principles that preserve the heterogeneity of the data. By integrating the historical censuses, we expect researchers to make greater use of the censuses again, now with 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. Although the harmonization of the data is still ongoing, we have already created methods and tools to provide a solution in RDF which is flexible enough to deal with changes and challenges of harmonizing aggregated data. We do this while not keeping the data in a self-contained environment, benevolently to stimulate use and inspire new links to the census.

NOTES

1. Dynamic register of the population per municipality
2. https://github.com/CEDAR-project/MP2Demo
3. http://www.w3.org/TR/rdf-primer/
4. http://www.w3.org/RDF/
5. http://www.w3.org/TR/rdf-sparql-query/
6. http://dbpedia.org/
7. http://datahub.io/dataset/2000-us-census-rdf
8. http://www.w3.org/TR/vocab-data-cube/
9. http://www.openannotation.org/spec/core/
10. https://github.com/Data2Semantics/TabLinker
11. http://lod.cedar-project.nl/cedar/sparql
12. http://lod.cedar-project.nl/cedar/sparql
13. http://www.cedar-project.nl/visualizing-sparql-query-results-on-the-census/
14. See http://lod.cedar-project.nl/cedar/sparql
15. See http://lod.cedar-project.nl/cedar/data.html
16. See http://www.cedar-project.nl/wp-content/uploads/datamodel.png
17. See http://lod.cedar-project.nl:8888/cedar/
18. See https://github.com/CEDAR-project/DataDump
19. See https://github.com/CEDAR-project

References

Begthol, C. 2010. Classification theory. *Encyclopedia of Library and Information Science* 3:1045–60.

Benford, F. 1938. The law of anomalous numbers. *Proceedings of the American Philosophical Society* 78 (4):551–72.

Berners-Lee, T., J. Hendler, and O. Lassila. 2001. The Semantic Web. *Scientific American* 284 (5):34–43.

Bukhari, A. C., and C. J. O. Baker. 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, July.

Den Dulk, K., and M. Van Jacques. 1999. The population censuses in the Netherlands. In *A century of statistics. Counting, accounting and recounting in the Netherlands*, ed. J. G. S. J. van Maarseveen and M. B. G. Gircour, 303–34. Voorburg, Amsterdam: Statistics Netherlands.

Doorn, P., J. Jonker, and T. Vreugdenhil. 2001. Digitalisering van de Nederlandse volkstellingen 1795–1971: Met een nadere beschouwing van de gedigitaliseerde telling van 1899. In *Nederland een eeuw geleden geteld: Een terugblik op de samenleving rond 1900*, ed. J.G.S.J. Maarseveen and P.K. Doorn, vol. 2, 41–64. Amsterdam: St. Beheer IISG.

Esteve, A., and M. Sobek. 2003. Challenges and methods of international census harmonization. *Historical Methods* 36 (2):37–41.

Fernández, J. D., M. A. M. Prieto, and C. Gutiérrez. 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)*, 20–25. New York: ACM.

Higgs, E. 1996. A clearer sense of the census: Victorian censuses and historical research. *Public Record Office Handbooks*, no. 28. London: Her Majesty's Stationery Office.

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

Meroño-Peñuela, A., A. Ashkpour, M. van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach, and F. van Harmelen. 2015. Semantic technologies for historical research: A survey. *Semantic Web – Interoperability, Usability, Applicability* 6 (6):1–27.

Meroño-Peñuela, A., A. Ashkpour, L. Rietveld, R. Hoekstra, and S. Schlobach. 2012. *Linked humanities data: The next frontier? A case-study in historical census data*. Proceedings of the 2nd International Workshop on Linked Science 2012 (LISC2012), ISWC 2012, Boston, MA.

Meyer, P. B., and A. M. Osborne. 2005. Proposed category system for 1960–2000 census occupations. Bureau of Labor Statistics Working Paper 383.

Muurling, S., and K. Mandemakers. 2012. MOSAIC census inventory of the Netherlands. Final report, (IISG Amsterdam), MOSAIC working paper WP2012-006. http://www.iisg.nl/hsn/documents/mosaic-wp-2012.pdf.

Petrou, I., and G. Papastefanatos. 2014. Publishing Greek Census Data as linked open data. *ERCIM News* 2014:96.

Pineo, P. C., J. Porter, H. A. McRoberts. 1977. The 1971 census and the socioeconomic classification of occupations. *Canadian Review of Sociology* 14 (1):91–102.

Putte, B. V. D., and A. Miles. 2005. A social classification scheme for historical occupational data. *Historical Methods* 38 (2):61–92.

Ruggles, S., and R. R. Menard. 1995. The Minnesota historical census projects. *Historical Methods* 28 (1):6–10.

Van Leeuwen, M. H. D., I. Maas, and A. Miles. 2002. *HISCO: Historical International Standard Classification of Occupations*. Leuven: Leuven University Press.

Van Maarseveen, J. 2008. *Dutch occupational censuses 1849–1971/2001. A component of the population census*. Voorburg, The Netherlands: Netherlands Central Bureau of Statistics.

Van Maarseveen, J., and P. K. Doorn. 2001. *Nederland een eeuw geleden geteld. Een terugblik op de samenleving rond 1900*. Amsterdam: Stichting Beheer IISG, 316.

Van der Meer, A., and O. Boonstra. 2006. *Repertorium van Nederlandse gemeenten 1812–2006*. Den Haag: Data Archiving and Networked Services.