Detailed Tables from the Dutch Census 1947:
Experiences and Lessons Learned in Publishing
a Large Dataset

Social and Economic History

Jan Jonker (corresponding author) | ORCID: 0000-0002-3014-5919
Data Archiving and Networked Services (DANS), Royal Netherlands Academy of Arts and Sciences (KNAW), The Hague, Netherlands
jan.jonker@dans.knaw.nl

Wouter Poot | ORCID: 0000-0002-6047-326X
DANS, KNAW, The Hague, Netherlands

Peter Doorn | ORCID: 0000-0002-8246-6757
DANS, KNAW, The Hague, Netherlands

Abstract

Since the end of the nineties, Dutch census publications have been digitized and made available for digital processing. New analyses of the data were presented in some fruitful conferences in the first decade of this century. In addition to the census publications, a mass of detailed census data was found in dossiers and so-called “transparencies” in the archive of Statistics Netherlands. Most of that material was scanned into digital images, awaiting further content conversion into numeric data. In the present article, the authors describe the process of digitizing the detailed tables of the Dutch Population and Occupational Censuses held in 1947, which is the first set of detailed census data that is made available in a digitally processible form. They give an example of historical analyses made possible by this dataset. Moreover, they take these census data as an example of preparing and publishing a large dataset. Experiences and lessons learned in the process lead to ample opportunities for further analysis of the data and for efficient ways to accomplish the content conversion of the many remaining images of census data.
Keywords


Online publication date: 26-10-2021


1 Introduction

Censuses are among the basic information sources on the state of a country. In the 1990s historians in many countries started initiatives to digitize historical censuses, most of which had been recorded and published on paper until well into the second half of the 20th century. In 1996 a proposal was formulated to digitize the printed publications of the Dutch Population censuses held from 1795 to 1971 (Doorn et al., 2001). In 1997 this led to a project of the Netherlands Historical Data Archive (now part of Data Archiving and Networked Services – DANS) in co-operation with the “Centraal Bureau voor de Statistiek” (CBS, Statistics Netherlands). As a result, on the centennial celebration of the CBS in 1999, the scanned images of the printed publications, about 42,500 pages in total, were published, originally on CD-ROMs. Additionally, of the census of 1899, the most voluminous work of the series, both the introductory volume and all published tables were converted into machine-readable form, published as a website (www.volkstelling1899.nl) and on the CBS Open Data Portal Statline. The digital census of 1899 gave rise to a series of new analyses, presented at a symposium and published in the book Nederland een eeuw geleden geteld. Een terugblik op de samenleving rond 1900 (Van Maarseveen & Doorn, 2001).

As a result of several projects run in the years 2002–2004, the data of all publications of the censuses from 1795 through 1971 were converted into a digitally processable form: pdf for the textual parts, Excel for the tables. This time a twenty-odd number of new analyses were presented at a symposium and published in book form by Boonstra et al. (2007). The first article in that monograph gives an overview of the preceding decade of digitizing the Dutch census publications (Doorn & van Maarseveen, 2007).

In 2004 the digitized data were made accessible for a worldwide audience via the website www.volkstellingen.nl and in the DANS data archive. The

In the present article, we describe the process of digitizing the detailed tables of the Dutch Population and Occupational Censuses held in 1947, the first census held after World War II. The publications of this census have been digitized in the above-mentioned projects. In addition to the publications, the CBS administers detailed data for 1947 in hand-written tables on so-called “transparencies”, about 30,000 sheets in A4 format. For later censuses, comparable detailed data already existed in digital form. We take the 1947 census as an example of preparing and publishing a large dataset. Our focus is on the challenges in digitizing such a volume of census data, in particular the processing of a large number of files.

Digitizing costs are one element to be reckoned with: costs for scanning and the high costs for converting the scanned images into electronically processable data. Transportation is another aspect. The transportation of tens of thousands of vulnerable originals, the transparencies, from the depot to the scanning location and back is a logistic operation, involving an accurate organization and registration of the originals as well as of the scanned images. Subsequently, the images should be stored and transferred. In every step, the large number of files requires a special, well-managed workflow.

2 Data

– Temporal coverage: 1947

The detailed tables of the Population Census 1947 are now available via the online archiving system of DANS. The data are presented in a collection of 12 datasets, one per province plus a dataset for the Netherlands as a whole. Each dataset contains data and documentation, divided into four folders: three containing the data and one containing the documentation. Figure 1 shows the structure of the dataset for the province of Drenthe.

The data consists of scans of the detailed tables in JPG format, XLS sheets containing the data of the detailed tables created by data entry, and CSV files, converted from the XLS sheets. The scans and the output files in XLS format are included as provenance data and to enable reproducibility. The output files in CSV-text format are best used for further analysis.
Within each dataset, the three data types have a separate folder. In this folder, the data is grouped by municipalities. Each dataset also contains documentation regarding the data: a description for the given dataset, a description of each detailed table, and the templates for the data entry.

To give an example, Figure 2 is the scan of a transparency for Table 4 (and Table 5) of the city Alkmaar in the province of North Holland. Figure 3 shows the Excel sheet resulting from the data entry of Table 4.

The list of tables of the 1947 census is, with minor variations, the same for each province. However, the degree of detail does vary within the provinces, where the larger cities are treated in more detail than the smaller villages. A list of tables provides the best overview of the richness of the source and the subjects treated (see Appendix).

The list makes clear that a great variety of detailed analyses of the Dutch population briefly after World War II is possible with the digitized data. Household composition, age structure, religious denomination, marital status, housing conditions, employment and work situation, nationalities,
commuting, and female participation in the labor force are just a selection of the topics that can be studied. Most of the information is available to a great geographic detail: not only for every one of the 1016 municipalities in 1947 but often for subdivisions of the municipality (neighborhoods, hamlets, etc.).
In section 4 of this article, we present one analysis of the dataset by way of example.

2.1 Special Feature of the Dataset
A special feature of this dataset is the presence of control counts in the output files. For any table that shows (sub)totals of rows and/or columns, the XLS files contain corresponding control rows and control columns. See, for example, Figure 3: if a check value is not equal to 0, it indicates an inconsistency in the table. That indicates either an error in the data entry or an error in the original material. A user of the dataset can find the cause of the inconsistency by comparing the original scan with the output files. The presence of the control counts is a valuable addition to other census datasets. Not only are control counts useful to end-users, but we also believe they should be part of census data by default. In this way, the data entry can be validated more easily, which benefits the scientific character of the dataset.

3 Methods
In this chapter we explain the way of creating the data, the extensive process of post-processing the data, and the steps followed for archiving and presenting the data.

3.1 Creating the Data
At the start of the project, about 30,000 transparent sheets in A4 format were available, containing detailed tables of the Dutch Population and Occupational Census held in 1947, shortly pc 1947. In 2005 these transparent sheets have been scanned into digital images.

The next step was to convert the images into digitally processible tables. In 2006 a pilot project aimed at performing this transformation by a smart form of Optical Character Recognition (OCR). However, at that time data entry appeared to be substantially less costly than OCR. Therefore, the choice has been made for full data entry (Jonker, 2008).

3.1.1 Organizing Data-Entry
To manage the data entry process efficiently, extensive instructions were formulated, including a technical description for the 33 table types that were distinguished. For each of the table types, one or more templates for Microsoft Excel were made, as illustrated in Figure 3. The templates in Excel form a
3.2  **Post-Processing**

The overall goal was to publish a FAIR dataset – Findable, Accessible, Interoperable, and Reusable. Publishing a dataset in EASY contributes a lot to the first three aspects of FAIRness. For post-processing the data, the focus was on Reusability. This requires transparent procedures to enable reproducible results with a clear meaning for later analysis. To promote transparency, the file names of the images represented the names of the respecting municipalities. Following the instructions, data entry should produce Excel files with filenames conforming to the names of the images. We have given much effort to checking and correcting that relation between images and Excel files.

3.2.1  Determining the Relation between Images and Output Files

The Excel files produced by data entry have been received by FTP. Ultimately, for many Excel files, two or more copies were received, counting to a total of 53,760 files. So, the first step was to select unique instances of the files. As it was impossible to inspect all files visually, the checks were done based on extensive file lists. For specific small subsets of files, the output files or image files were inspected visually. The names of all output files of the data entry were checked against the image names. Mistakes have been corrected.

3.2.2  Determining the Relation between Images and Tables

The image files for each municipality were numbered consecutively. However, for some table types, the number of images varies with the size of the municipality. Consequently, it was necessary to determine which table type is described in which image file and its corresponding output file(s).

Because every sheet has a header containing the name of the municipality and the name of the neighborhood, if applicable, the clue was in analyzing those headers. We have done that by converting the Excel files to CSV-text files. The large number of files made it necessary to use an automatic procedure, for which we used the utility Excel Converter (2017).

Using some MS-batch scripts, defining appropriate labels for the table types and after a few minor manual corrections, the relation between output files, the corresponding images, and table types was described. As a result, we created 33,034 unique output files, which correspond to 28,765 image files.
3.2.3 Assigning Version Numbers to Output Files

The final step of post-processing was to assign version numbers to the output files. The 33,034 unique output files received the version number V0 if they were unique for the corresponding image. Otherwise, they received version number V1. For the V1 files, companion files got version numbers V2, etc. For end-users of the dataset, only the V0 and V1 versions are accessible. V1 files indicate that an earlier version is available at DANS if users would want to check the version history. The output files are consistently available in Excel and in csv format.

3.2.4 Post-Processing – a Procedure and a Method of Documentation

The process of post-processing illustrates the tedious attempt to get grip on this large dataset. By repeated sorting, counting, and comparison between extensive file lists, we searched for unexpected patterns in the sets of files. If we found any “inconsistencies”, they have been corrected. The results of the corrections were checked by a systematic comparison of file lists “before” and “after” correction.

Reproducibility of such a process requires adequate documentation, of course. The overall process is described in a series of tables (Excel workbooks) with accompanying documentation. The documentation describes the composition of the tables in the Excel workbooks step by step, worksheet by worksheet. Corrections concerning a moderate number of files were executed interactively and documented in the relevant Excel workbook. Corrections concerning a larger number of files were executed by way of MS-batch scripts and documented internally in the scripts. The documentation of all this is contained in some 60 Excel worksheets and more than 60 auxiliary documentation files.

3.3 Archiving and Presenting

Archiving and presenting census data presents several challenges. We address three of them. Firstly, the quest for IT tools capable of handling a large number of files. Secondly, designing the final data structure in the digital repository itself, the DANS archive. Thirdly, the process of uploading the datafiles and the metadata to the digital repository.

3.3.1 Choosing IT Tools

As mentioned before, the main problem in dealing with the detailed census data of 1947 was the considerable number of files. Therefore, applications must be used that execute fast. Moreover, they must be easily transferable between people, remaining affordable, and without a need for extensive training. In
addition, the applications must work transparently, to enable correct interpretation of the results of each operation. Finally, the application must support proper documentation of the operations executed.

Given these requirements, we chose for MS Office (Excel with corresponding documentation in Word), MS-batch-scripts (home-made) and the Excel Converter utility by Svelte (2017). In the phase of uploading files to the DANS archive we used ASAP Utilities (2018) to convert accented characters, as described below.

3.3.2 Designing the Final Structure of the Dataset
When designing the data structure for the present dataset, the focus was on dissemination. For this purpose, the dataset was divided into 12 subsets, one dataset for each province of the Netherlands, plus one for the country as a whole. Using this structure, the number of files of the largest subset was brought down to about 17,000. The numbers of data files of this order of magnitude are a challenge for many repository systems, including EASY.

As the time needed to display a dataset in EASY is proportional to its size, lowering the number of files per dataset provides a significant advantage to end-users. An even lower number of files per dataset could be reached by creating a subset per municipality, but as the Netherlands had more than a thousand municipalities in 1947, this method would not be feasible.

In addition to the improved data structure, this version of PC 1947 has another distinct advantage: individual file metadata. For each image file, the corresponding table type is added to the metadata of the scan. End users can directly see this relation in the user interface of EASY. When downloading the data, a spreadsheet is available with that metadata for all files in the selected dataset.

3.3.3 Uploading Datafiles and Metadata to the Digital Repository
After the post-processing was completed and the data structure of the dataset had been designed, the data were prepared for uploading to the digital repository. First, the data for each subset per province was grouped into folders for the scans and the output files, respectively. Each dataset also contains a folder with a descriptive document and additional metadata regarding the table structures, the relation between scans and tables, and the templates used for data entry. Next, the data was uploaded to the repository. We did not run into major difficulties when uploading the data itself, despite the large number of files. In a third step, the metadata was uploaded. During this step, however, several errors did occur. It turned out that EASY does not accept accented characters in its XML-based metadata, so we had to replace accented characters with...
their unaccented counterparts. This time, uploading the metadata succeeded without any issues and all subsets could be published.

4 Using the Dataset

The goal of digitizing data collections is to make the data reusable. In this section, we give an example of using the dataset of the Dutch Population Census 1947 for historical analysis. Next, we offer a perspective for several ways to broaden the scope for re-using the dataset.

4.1 Using the Dataset for Historical Analysis

As indicated above in section 2, the detailed dataset offers extensive possibilities for statistical historical research about the situation in the Netherlands after the end of World War II. As an example, we consider the part of the population that did not have the Dutch nationality. The year 1947 is well before substantive international migration streams, both as a consequence of decolonization (from Indonesia, the former Dutch Indies; and later from Surinam and the Antilles), economic needs (guest labor), and political conditions (refugees, asylum seekers) elsewhere, began.

In part A1 of the Census publications, Table 5 shows a relatively large part of Germans and Belgians, but also quite some people without nationality or with nationality unknown. The latter categories comprise people who had lost their nationality during the war period and were waiting for the treatment of their application for Dutch nationality. Because of this complication, we focus here on the total of people without Dutch nationality in relation to the total population per municipality.

For the analysis, we used the csv-text files per municipality for Table 7 (Bevolking naar werkkring en nationaliteit), focusing on nationality. The calculation required several steps: first, we checked whether the files contained inconsistencies. By summarizing the lines with checksums, we noted that one file (Brouwershaven_003C_v0.csv) missed a column total, which could however easily be fixed.

Next, we combined the files for Table 7 per municipality into one comprehensive file. Subsequently, we computed the ratio of people without Dutch nationality per 10,000 of the population. Finally, we combined the results in a file with vector coordinates (Boonstra, 2007) to produce Figure 4, a map that shows the relative distribution of the non-Dutch population in the Netherlands in 1947.

The map is made with the QGIS application, which automatically generated categories with about 1/6 number of the municipalities. We see that
Figure 4  Persons without Dutch nationality per 10,000 of the population, by municipality 1947
relatively many of the non-Dutch people lived in border regions, but also in some Western parts of the Netherlands. The urban ring ‘Randstad’, a term originally coined around 1930 by the founder of KLM (Royal Dutch Airlines), Albert Plesman, is already discernible on the map.

4.2 Perspectives to Expand Usage of the Dataset

Several features of the dataset support further analyses: the extensive documentation, the presence of control counts, and last but not least the availability of the output in tabular csv-text format. We mention three directions for further usage of the dataset.

a. In the previous section, we mentioned two preparatory steps before we computed the ratios of people without Dutch nationality. First, we checked for inconsistencies in the tables. Next, we combined the files per municipality into one comprehensive file for the table considered. To enhance the reusability of the data, one could perform these steps in advance and publish the results.

b. Another way to reuse the data is from the perspective of linked data. The analytic potential of linked data created from the published census tables is described in Meroño-Peñuela (2016), Meroño-Peñuela et al. (2017), and Ashkpour (2019). The organization of the dataset of the Population Census 1947 enables efficient preparation for the conversion to RDF for linking the data to other censuses and related population data.

c. In addition to analyzing the rich content of the dataset of the 1947 census, the dataset gives a possibility to evaluate the potential of advanced OCR tools as an alternative to data entry. Applying new OCR methods to selected images in the present dataset will enable validation of those methods. Subsequently, the positively validated OCR tools could be used for the content conversion of the many other scanned images of detailed tabular data that are available in the DANS repository.

5 Concluding Remarks

5.1 Lessons Learned

The analysis of a really large number of data files, largely without visual inspection of the files, requires a special indirect method of working and an adequate way of documenting. We developed a way of processing the data and a way of documenting the process that is transferable and reproducible.

We used standard tools from the MS Office suite for a large part of the work, plus basic programming utilities like MS-batch scripts. Additionally, we used
web applications for specialized actions such as the conversion of Excel files to CSV-text files. The tools used are accessible to most users, do not require a steep learning curve and, most of all, were suitable for the treatment of large quantities of files. In retrospect, the way we organized the data entry has been successful. However, the synchronization of filenames appeared to be error-prone and could better be taken care of prior to data entry.

Given the time it took for the whole operation, some all-time favorite practices proved to be productive: making files read-only at the end of the day and distinguishing versions by starting filenames with a date indication, for instance, “yyyyymmdd”.

5.2 Conclusion

Historians and social scientists use many different types of sources to help them create an image of society, contemporary and in the past. Historical censuses are a special category, as they allow for the detailed study of the population over time. Our dataset with detailed tables of the Population Census 1947 forms such a source for the Netherlands shortly after World War II.

In retrospect, we think that this dataset is valuable in several ways. As we illustrated in section 4.1, the dataset enables historical analyses on a detailed level of municipalities or even on the sub-municipal level. Further, by providing extensive documentation and presenting the datafiles in appropriate formats, the dataset enables (preparation for) applied analysis in other environments, for example in the context of linked data.

In addition to using it in historical statistical analysis, the dataset and the accompanying documentation on its content and provenance offer a feasible starting point for the content conversion of other large sets of images concerning tabular data.

In conclusion, the detailed tables from the “Population and Occupational Censuses 1947” represent a valuable example of a dataset that is FAIR — Findable, Accessible, Interoperable, and Reusable.

References

[All hyperlinks last accessed on 2021-10-04]


Appendix

List of Tables of the Dataset for the 1947 Census

In section 2, we refer to the list of tables contained in the dataset. Figure 1 shows the document Titles of Tables as a part of the documentation in the dataset. However, that document is in Dutch. Table A1 contains the titles of the tables in English translation.

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Population in each part of the municipality (by type of inhabited places)</td>
</tr>
<tr>
<td>Table 2</td>
<td>Population by position in the household</td>
</tr>
<tr>
<td>Table 2a</td>
<td>Population aged 65 and older by position in the household and year of birth</td>
</tr>
<tr>
<td>Table 3</td>
<td>Population by year of birth</td>
</tr>
<tr>
<td>Table 4</td>
<td>Population by age group and marital status</td>
</tr>
<tr>
<td>Table 4a</td>
<td>Population by year of birth and marital status</td>
</tr>
<tr>
<td>Table 5</td>
<td>Population by age group and type of living quarters</td>
</tr>
<tr>
<td>Table 6a</td>
<td>Population by age groups and main denominations</td>
</tr>
<tr>
<td>Table 6b</td>
<td>Some smaller denominations (included in columns 20 to 23 of Table 6a)</td>
</tr>
<tr>
<td>Table 6c</td>
<td>Population by year of birth and denomination</td>
</tr>
<tr>
<td>Table 7</td>
<td>Population by employment and nationality</td>
</tr>
<tr>
<td>Table 8</td>
<td>Population by place of birth and year of birth classes</td>
</tr>
<tr>
<td>Table 9</td>
<td>Population by current employment and place of residence in August 1939</td>
</tr>
<tr>
<td>Table 10</td>
<td>Working population by classes of economic activity and employment status</td>
</tr>
<tr>
<td>Table 10a</td>
<td>Working population (excl. temporarily not working) by groups of economic activity and employment status</td>
</tr>
<tr>
<td>Table 10b</td>
<td>Working population by classes of economic activity and employment status, per neighborhood</td>
</tr>
<tr>
<td>Table 11</td>
<td>Working population by classes of economic activity and age classes</td>
</tr>
<tr>
<td>Table 11a</td>
<td>Working population by classes of economic activity and age classes, per neighborhood</td>
</tr>
<tr>
<td>Table</td>
<td>Title</td>
</tr>
<tr>
<td>----------</td>
<td>-----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Table 12</td>
<td>Working population by classes of economic activity and position in the household</td>
</tr>
<tr>
<td>Table 13</td>
<td>Persons working in government service (excluding temporary military service) by classes of economic activity</td>
</tr>
<tr>
<td>Table 14</td>
<td>Working population (excl. temporarily not working), working in or outside the residential municipality, by mode of transport, by class of economic activity</td>
</tr>
<tr>
<td>Table 14a</td>
<td>Working population (excl. temporarily not working), working in or outside the residential municipality, by mode of transport, by class of economic activity, per neighborhood</td>
</tr>
<tr>
<td>Table 14b</td>
<td>Commuters by municipality of work, mode of transport, employment, employment status and position in the household</td>
</tr>
<tr>
<td>Table 15</td>
<td>Working population (excl. temporarily not working) by occupational groups and age classes</td>
</tr>
<tr>
<td>Table 15a</td>
<td>Workers, employees, and persons in the liberal professions, by main occupations</td>
</tr>
<tr>
<td>Table 15b</td>
<td>Working population (excl. temporarily not working) by occupational groups and denomination</td>
</tr>
<tr>
<td>Table 15c</td>
<td>Working population (excl. temporarily not working) by occupational groups and nationality</td>
</tr>
<tr>
<td>Table 15d</td>
<td>Workers, employees, and persons in the liberal professions by main occupations and age classes</td>
</tr>
<tr>
<td>Table 15e</td>
<td>Workers, employees, and persons in the liberal professions, by main occupations and marital status</td>
</tr>
<tr>
<td>Table 15f</td>
<td>Male heads of families with occupation (excl. temporarily not working) by occupational groups and denomination</td>
</tr>
<tr>
<td>Table 15g</td>
<td>Male heads of households with occupation (excl. temporarily not working) by occupational groups and classes of economic activity</td>
</tr>
<tr>
<td>Table 16</td>
<td>Population of the neighborhoods</td>
</tr>
<tr>
<td>Table 17</td>
<td>Work commuters by municipality of residence, mode of transport, employment, employment status and position in the household</td>
</tr>
</tbody>
</table>