The CEDAR Project: Publishing and Consuming Harmonized Census Data

Albert Meroño-Peñuela[1], Ashkan Ashkpour[2], Christophe Guéret[1] and Andrea Scharnhorst[1]
[1]Data Archiving and Networked Services (DANS), Den Haag, NL
[2]International Institute of Social History, Amsterdam, NL
{albert.merono, christophe.gueret, andrea.scharnhorst}@dans.knaw.nl
ashkan.ashkpour@iisg.nl

Abstract. This paper discusses the use of semantic technologies to increase quality, machine-processability, format translatability and cross-querying of complex tabular datasets often found in many research areas of the Humanities. In particular, we are interested in enabling longitudinal studies of social processes in the past. We use the historical Dutch censuses as case-study: census data is currently digitized, but it is notoriously difficult to compare, aggregate and query in a uniform fashion. We describe an approach to achieve these goals, emphasizing open problems and trade-offs.

1. Introduction

Census data plays an invaluable role in the historical study of society. In the Netherlands, the Dutch historical censuses (1795-1971) are among the most frequently consulted sources of statistics of the Central Bureau of Statistics[1] [1]. During the period they cover, it is the only regular statistical population study performed by the Dutch government, and the only historical data on population characteristics that is not strongly distorted. Over the past decades there have been a lot of digitization efforts across the world in bringing (historical) census data to researchers. Historical censuses comprise very detailed data about specific categories and variables in a certain time period in history, which is why historians are greatly interested in the digitizing of historical census data. The censuses are a rich source of historical information for researchers providing demographic, social and economic structures, yielding a wealth of data on many issues in the course of time [4].

The dataset comprises 507 Excel workbooks with 2,288 tables, being a digital representation of a partial subset of the original census books, which only contain aggregated data. These books have been translated to several digital formats in subsequent stages: first as scanned images, then as PDF documents, and finally as Excel tabular files. The dataset is archived at DANS' archiving system EASY[2], and publicly available at www.volkstellingen.nl as open data.

[1] http://www.cbs.nl
[2] http://easy.dans.knaw.nl

Nevertheless, the usability of the dataset by citizens and researchers is a daunting task. The tabular files are not well self-described, and users can spend hours of manual seeking before finding the data they are interested in. Moreover, the richness of variables in the dataset (e.g. age, gender, religions, occupations, locations, housing types...) together with drifting variables as a consequence of time, make automated longitudinal queries impracticable in tabular format. Without these cross queries and a systematic access to fine-grained information, stories from data stay away from citizens and researchers.

The project CEDAR[3] from the eHumanities group will enable a smoother access to the historical census data. In the following, we expose the issues we have encountered and the implementations we have developed with respect to format transformations, quality checking and combination of this dataset with others, as well as describing open problems still pendant to solve.

1.1 Harmonization Challenge

The harmonization of the Dutch historical census data highlights the problems which historians deal with, but also other digital humanities researchers face when working heterogeneous datasets. The nature of such a dataset necessitates different harmonization approaches. Next to the changing structures, evolving classifications and drifting concepts, the digitized Dutch census provides many practical harmonization challenges where we need more than linking and statistical methods. In many cases, information about the lower levels of the municipalities has merged with other columns in the Excel files, making it difficult to extract this type of information. Due to the size of the Excel tables, which often have more than ten thousand rows, it is unpractical to restructure the data by hand. The practical challenge here is to use computational techniques in order to clean the data before continuing with the harmonization. We define census harmonization as a bottom-up multilevel approach where we discuss a set of census specific harmonization methodologies in order to deal with the different types of information contained in the census and its changing structure over time.

2. Census Data Open Linked

The CEDAR project aims at exposing census open data on the Web, making it more accessible, linkable and queryable [2]. In CEDAR we apply a specific web-based data-model (exploiting the Resource Description Framework - RDF - technology) to make census data inter-linkable with other hubs of historical socio-economic and demographic data and beyond. The project will result in generic methods and tools to weave historical and socio-economic datasets into an interlinked semantic data-web.

2.1. Complex Tabular Data to RDF

The census tables are exceptionally messy in their layout and contents. Structural heterogeneity is present in row and column headers that do not follow any recognizable pattern between different tables. Sometimes contents of row and column headers are defined hierarchically,

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

spanning through several columns or rows. Other issues, like unstructured subtotals in the data cells, notes written outside of the table borders, and unstructured provenance in annotations make it even harder to identify a common structure. Additionally, semantic heterogeneity makes the meaning and value of variables incompatible in cross-year comparisons. Traditionally, social historians have solved these problems with *harmonization*, a complex process of data restructuring and alignment performed, until now, by hand.

To gain access to a more fine-grained atoms of these tabular data, we converted the dataset to RDF with our own tool, TabLinker[4]. TabLinker produces an RDF named graph per census table according with the RDF Data Cube Vocabulary[5]. Other csv2rdf scripts were discarded due to the high complexity of the census tables. TabLinker requires prior table formatting with styles that define specific cell bounding boxes.

2.2. Images, PDF, Tables, RDF Data Cube: Formats on-demand

One of the CEDAR premises is to provide data to stakeholders in their favourite formats. Despite our take for RDF in solving harmonization in an automated manner, data should be delivered as requested on-demand by users, including tabular formats, structured formats like XML or JSON, RDF or relational databases (according to some relational schema). Checking whether we can round-trip among these formats is one of our subjects of study.

Additionally, we are aware that some tasks final users, especially researchers, may want to do require not only accessing the data atoms, but also referencing the books, images, PDFs and Excel tables these data come from. In other words: extensive provenance about the data source and the transformations these data may have suffered.

2.3. Quality Assessment: Annotations and Cleaning

The census tables contain 33,283 annotations made by the original census makers, digitizers and social historians. The contents of these annotations are varied: some are descriptive and provide additional details about meaning of variables or constraints of values, while others contain proper corrections on the data, like cells that do not sum up, or individuals that are considered exceptions and are counted into an incoherent category.

TabLinker also produces independent RDF graphs for these annotations, using the Open Annotation Data Model[6]. Despite of representing legacy annotated knowledge, the possibility of creating new annotations provides a useful mechanism to receive feedback on the dataset, enabling a correction workflow that improves data quality without harming the original values. This way, the dataset can be exposed online, allowing experts to input their own annotations. However, when final users perform queries these annotations need to be harvested to provide appropriate responses. The way in which annotation's contents need to be compared, selected

[4] https://github.com/Data2Semantics/TabLinker
[5] http://www.w3.org/TR/vocab-data-cube/
[6] http://www.openannotation.org/spec/core/

or transformed for that purpose is no trivial task.

The study of the content of some of these annotations leads to consider the dataset quality. How can we be sure about the correctness of all the values on these cells? To answer this question, we developed a script[7] that checks Benford's Law[8] for the census dataset. This is a very well known rule that censuses have historically met, and the script proves that this is also the case for the Dutch historical censuses. We provide these results as a measure of the quality of the data, and we claim that proof of data quality has to be always provided with open data. Because of that, we generate dataset metadata with additional tools[9], and we are working on how to integrate data quality measures in dataset self-descriptions.

2.4. Longitudinal Queries

Representing tabular data into RDF does not solve the problem of cross querying *per se*. Solving longitudinal queries requires combining the census data with other datasets. For instance, answering a cross query about occupational titles (e.g. *evolution on the number of shoemakers in Amsterdam from 1795 to 1971*) needs a mapping between all cells of all tables that describe properties of shoemakers with the class *Shoemaker* of the Historical International Standard Classification of Occupations[10] (HISCO), that represents all individuals with that historical occupation. Then, that class can be queried explicitly to retrieve shoemakers, no matter the way they are spelled in a specific table.

In CEDAR we need to combine a variety of different data sources to solve harmonized queries. If these data sources come in a non-RDF format (and they usually do), we rely on our conversion scripts (like TabLinker) to express them in standard vocabularies that make them suitable for linking. If no such standard vocabularies exist, we arise the need for creating and standardizing them.

3. Knowledge discovery, ontologies and concept drift

Solving longitudinal queries also requires to extract some inner knowledge representations that are implicit in the census tables. For instance, occupations follow a tree-like structure, where all labour types in one year are classified, from very generic classes (e.g. *Class I*, *Class II*), to very specific ones (e.g. *shoemakers that own a shoe company*). We use the graphs extracted by TabLinker and, by ways of SPARQL[11] queries, we extract the trees following a bottom-up approach.

Each census year is characterized by one of these trees or ontologies, because the occupation classification system changed for each census. It is clear, then, that these different ontologies

---

[7] https://github.com/CEDAR-project/MP2Demo
[8] http://en.wikipedia.org/wiki/Benford's_law
[9] https://github.com/CEDAR-project/TabExtractor
[10] http://historyofwork.iisg.nl/

[11] http://www.w3.org/TR/rdf-sparql-query/

need to be aligned in order to solve longitudinal queries. We need to establish a mapping between the classes that defined a labour type across all these ontologies; otherwise, since the labels and number distributions change over time, their meaning is isolated and unrelated.

A specific problem we encounter in this scenario is concept drift [3]. While mapping concepts from different times, it may happen that one of these concepts has evolved its meaning. Concepts may drift in three different ways: intension drift (a change in the fundamental properties that define the concept), extension drift (a change in the individuals that belong to the concept), and label drift (a change in the linguistic terms that label the concept).

4. Conclusion

We have presented the CEDAR project and motivated a framework for effective machine-processing of disparate sources, a common situation in Humanities data. We have discussed a generic architecture that deals with complex table conversion, data quality, provenance, and longitudinal queries. We have discussed an in-out data transformation approach, transforming tabular data into standard RDF data models and providing results back in the format preferred by the user. We have shown a practical application of data combination to solve cross queries in RDF space. Finally, we have described our knowledge discovery approach and motivated the problem of concept drift.

References

[1] Boonstra, O., et al.: Twee eeuwen Nederland geteld. Onderzoek met de digitale Volks-, Beroeps- en Woningtellingen 17952001. DANS en CBS (2007)

[2] Meroño-Peñuela, A., Ashkpour, A., Rietveld, L., Hoekstra, R., Schlobach, S.: Linked humanities data: The next frontier? a case-study in historical census data. Proceedings of the 2nd International Workshop on Linked Science (LISC2012). International Semantic Web Conference (ISWC) (2012), http://ceur-ws.org/Vol-951/

[3] Wang, S., Schlobach, S., Klein, M.C.A.: What is concept drift and how to measure it? In: Cimiano, P., Pinto, H.S. (eds.) EKAW. Lecture Notes in Computer Science, vol. 6317, pp. 241–256. Springer (2010)

[4] Van Maarseveen, J.,: Dutch Occupational Censuses 1849-1971/2001. A component of the Population Census. Netherlands Central Bureau of Statistics (2008)