Longitudinal Queries over Linked Census Data

Albert Meroño-Peñuela[1,2], Rinke Hoekstra[1], Andrea Scharnhorst[2], Christophe Guéret[2], and Ashkan Ashkpour[3]

[1] Department of Computer Science, VU University Amsterdam, NL
albert.merono@vu.nl
[2] Data Archiving and Networked Services, KNAW, NL
[3] International Institute of Social History, Amsterdam, NL

Abstract. This paper discusses the use of semantic technologies to increase quality, machine-processability, format translatability and cross-querying of complex tabular datasets. Our interest is to enable longitudinal studies of social processes in the past, and we use the historical Dutch censuses as case-study. Census data is notoriously difficult to compare, aggregate and query in a uniform fashion. We describe an approach to achieve this, discussing results, trade-offs and open problems.

1 Introduction

Census data have great value in the historical study of society. In the Netherlands, the Dutch historical censuses (1795-1971) are among the most frequently consulted statistic sources of the Central Bureau of Statistics [1]. During the period they cover, it is the only regular statistical population study performed by the Dutch government, and the only not strongly distorted historical dataset on population characteristics. Various efforts have been taken to make these data digitally available. Digitized images aside, the original censuses have been manually entered into 507 Excel workbooks and 2,288 tables.

The dataset is very difficult to use in its current form. The tables are not well self-described, and researchers must spend hours of manual analysis. The variety of variables in the dataset, on the one hand, and concept drift [3] through time (i.e. changes in the meaning of concepts, e.g. shoemakers evolved from *leather workers* to *company owners*), on the other hand, make automated, longitudinal analyses impracticable. We discuss issues, implementations and open problems on format transformations, quality checking and dataset interlinking.

2 Longitudinal Linked Census Data

Our approach is based on the Linked Data paradigm. We apply a set of standard vocabularies to make census data queryable and inter-linkable with other hubs of historical socio-economic and demographic data.

We translate the dataset to RDF with a supervised XLS/CSV to RDF converter, TabLinker[4]. TabLinker produces an RDF named graph per census table

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

expressed using the Data Cube Vocabulary [2]. This is a semi-automatic process: first experts need to mark manually the tables, and then TabLinker produces the T/A-Box according to table styles and data. Secondly, the census dataset contains 33,283 annotations attached to original values with data corrections, descriptions and interpretations that we translate to RDF using the Open Annotation Data Model. Query responses are provided in multiple formats (tabular, tree-structured, graph-structured, relational) as requested on demand by users.

However, the straightforward conversion of census tables into RDF does not solve cross querying across multiple tables *per se*. A valid longitudinal query would be, e.g., *get the evolution on the number of shoemakers in Amsterdam from 1795 to 1971*. But what counts as a *shoemaker* varies in meaning and labeling in different tables. Solving these queries requires additional, more abstract knowledge on these variations. We propose a two-fold approach to integrate this knowledge. First, as a top-down strategy, we transform existing taxonomies like the Historical International Standard Classification of Occupations (HISCO) into RDF/SKOS, and map disparate census table values with similar meaning into this general scheme. Second, as a bottom-up approach, we consider the specific local T-Boxes created by TabLinker for each table and we gradually generalize them. We do this by ways of mapping (e.g. using Silk), and manually creating upper ontologies with more abstract entities. Our evaluation relies on replicating results of previous socio-historical work on the census [1].

3 Conclusion

In this paper we present an overview of our efforts to improve quality, machine-processability, format translatability and cross-querying of complex historical tabular datasets. We advocate a supervised conversion tool to translate these datasets into RDF. We illustrate an hybrid top-down/bottom-up approach to build multi-layered ontologies, allowing us to query data longitudinally.

Acknowledgements The work on which this paper is based has been partly supported by the Computational Humanities Programme of the Royal Netherlands Academy of Arts and Sciences, under the auspices of the CEDAR[5] project. For further information, see `http://ehumanities.nl`.

References

1. Boonstra, O., et al.: Twee eeuwen Nederland geteld. Onderzoek met de digitale Volks-, Beroeps- en Woningtellingen 1795–2001. 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)

[5] `http://www.cedar-project.nl/`