From Napoleon Conquests to the Big Brother Sabotage: Harmonization of the Dutch Historical Censuses in the Semantic Web

Albert Meroño-Peñuela[1,2] and Ashkan Ashkpour[3,4]

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

Abstract. Around the turn of the 18th century, the first integral population enumeration was held in the Netherlands during the Batavian Republic. It took over 30 years before the first official census was, by royal decree, organized and conducted in 1829, and was meant to be held from then onwards every ten years. The Dutch historical censuses are the only large scale, reliable statistical datasets available about the (demographic, social and economic) history of the Netherlands, covering an all-encompassing geographical area for over two centuries (1795–1971). Not surprisingly, the currently preserved and digitized historical censuses are the most consulted historical statistics by researchers. However, the 2 288 census tables are highly disconnected and scarcely integrated in their current form. Meaningful information is still hidden in these missing table-links, meaning that this wealth of information is not reaped to its full potential. In this paper we describe the lessons learnt in CEDAR[5], a project of the Computational Humanities Programme[6], to provide solutions to these integration problems. Our system leverages semantic technologies and Linked Data practices, which allow us to convert the census tables into a graph of fine-grained Linked Census Data. Using the distributed architecture of the Web, we interlink this graph with other online historical socioeconomic and demographic Linked Datasets. We use the information provided by these external links to guide the harmonization process in our dataset. At the same time, we investigate which historical classifications are not online yet following Web standards, and we use our census tables (on demographic structures, housing types, occupational classes and statuses, and religious denominations) to urge the need of publishing these historical classifications on the Web. Such historical hubs could increase enormously the interoperability of other datasets. Finally, we propose a querying pipeline on the resulting harmonized census dataset to enhance the data exploration work by historians and social scientists and help answering their research questions.

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

[6] See http:///www.ehumanities.nl/