CEDAR: Harmonization of Historical Dutch Census Data

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

[1] International Institute of Social History, Amsterdam, NL
[2] Erasmus University, Rotterdam, NL
[3] VU University Amsterdam, NL
[4] Data Archiving and Networking Services, Den Haag, NL

10th European Social Science History Conference Vienna

Abstract:

Historical censuses comprise very detailed information about our past. The complex historic situation is coded using specific categories and variables, defined for a certain time period in history and carrying a specific meaning. The census is one of our few reliable and large scale historical statistical data sources which are not strongly distorted. In the Netherlands, historical census data has been collected in aggregated form for the period 1795-1971. The digitization of those sources has been undertaken in the last decade.

An important characteristic of historical censuses is their potential for the study of social and economic change over time. Historical censuses are however difficult to use in a systematic and longitudinal way. Census questions and outcomes were reorganized, using different structures in almost every census. These changes often reflect the information needed for government activities at specific times, but present serious problems for researchers studying categories and variables over longer periods of time. Thousands of heterogeneous unlinked tables which use different schemas over time make it a cumbersome task for researchers to use these data. Most studies using the census are therefore constrained to use only a fraction of the data (i.e. specific years).

This paper focuses on the harmonization of the Dutch historical census data. In an ongoing effort to provide better access and use to historical censuses, these data have been digitized from 1997 onwards and are now available as +2300 disconnected Excel sheets. The Dutch census data are currently still a relatively untapped source of information and provide many insights into the social history of the Netherlands, which have not been exploited yet when looking at the great potential of a harmonized census dataset.

The aggregated nature of the Dutch Census data necessitates different harmonization approaches compared to the current practices applied to micro data. Extant literature currently does not provide enough insights in the practice of data harmonization when dealing with aggregated census data. The

lack of comprehension into the workflow and harmonization of historical Census data is still a bottleneck for researchers.

The CEDAR (Census Data Research) project aims to offer greater access and ease of use to the Dutch historical censuses by reducing the sheer number of scattered tables and provide a harmonized database which allows different views over the data. In this paper we identify and define census harmonization as a multilevel approach, consisting of a set of specific harmonization methodologies. We show how these techniques help us to deal with the different types of information contained in the census and its changing structure over time. We describe the challenges and practice of historical census harmonization and define it through a bottom up approach.