An Ontology of Finnish Historical Occupations

Lia Gasbarra[1], Mikko Koho[1], Ilkka Jokipii[3,4], Heikki Rantala[1], and Eero Hyvönen[1,2]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
[3] The National Archives of Finland
[4] Faculty of Arts, University of Helsinki, Finland

Abstract Historical datasets often impose the need to study groups of people based on occupation or social status. This paper presents first results in creating an ontology of historical Finnish occupations, AMMO, that enables selection of groups of people based on their occupation, occupational groups, or socioeconomic class. For interoperability, AMMO is linked to the HISCO international historical occupation classification and to a late 20th century Finnish occupational classification. AMMO will be used as a component in two semantic portals for Finnish war history.

1 Introduction

The Finnish Civil War and the Second World War (WW2) are of great interest to historians and to the wider public: most Finns have relatives who died or participated in these wars as soldiers. For example, the WarSampo semantic portal [1] has had more than 400 000 users since its opening in 2015.

An interesting piece of information in war-related datasets are the occupations and social statuses of the soldiers. For example, historians studying the Finnish Civil War are interested in learning why some people survived the prisoner camps and some did not—is this related to the occupation and social status of the prisoners? The problem is, however, that there are circa 1400 occupational labels in the Civil War data alone without any structure.

This paper presents the work done so far in studying, and harmonizing Finnish historical occupational labels, consisting mainly of manual expert work. AMMO, our historical Finnish occupation ontology under development, will be a fundamental resource containing Finnish historical occupational labels from 1914–1945, based on the War Victims[5] and WarSampo[6] datasets, in SKOS format. AMMO occupations are linked to the international historical classification of occupations HISCO [7] that is based on historical data collections from European and North American economies. HISCO provides a hierarchical backbone of occupational groups, as well as social stratification information through several measures like HISCLASS and HISCAM [3]. AMMO is also aligned with the

[5] http://www.ldf.fi/dataset/narc-sotasurmat1914-22
[6] http://www.ldf.fi/dataset/warsa

Finnish Classification of occupations 1980 [6], a social stratification classification system in use in Finland.

This paper presents details on the manual harmonization work of national historical occupational labels in heterogeneous datasets. The AMMO ontology model and more detailed ontology creation process is presented in [2].

Related Work Belgian occupational titles have been linked to early HISCO, with an overview of the common coding problems [4]. The Historical Sample of the Netherlands contains a representative sample of 78,000 people born in the Netherlands during 1812-1922, where occupational titles have been standardized and mapped to HISCO [8]. HISCO has been applied to the historical data of Catalonian population from the fifteenth to the twentieth century [5].

2 Occupations in Historical Datasets

The source datasets of AMMO are presented in Table 1, containing personal information of a total of 139 091 historical persons (soldiers) annotated with thousands of different occupation labels. The amount of people per dataset, and distinct occupational labels are shown. In the datasets, alternative or abridged forms of the same occupational title are often present (for example: 'hitsari' and 'hitsaaja' for welder). In some cases, the occupational title of a person is actually not an occupation but a social role or status, such as 'student', 'nobleman', 'child', or 'tenant'. Although occupations in HISCO are by definition solely activities that generate a remuneration, it is also possible to categorize these social roles or statuses (e.g., student) through HISCO relation and status coding.

The First World War (WW1) War Victims dataset contains the deceased of the WW1 era, which consist mostly of the Finnish Civil War. The data is gathered originally from various heterogeneous Finnish sources. It is not solely based on primary sources, but is a result of research, interpretation, and unification of nonhomogeneous data. The database presents the interpretation of sources provided by their compilers during different phases of archival work. These entries were gathered in many cases from parish records taken at the moment of death and/or other registers of parties, unions, local administrations and alike.

Both of the WW2 datasets shown in Table 1 are part of the WarSampo data. The Prisoners of War dataset contains the actual information source for most of the occupation annotations, whereas the death records lack this information. The WarSampo persons present a balanced sample of the Finnish male population during WW2, as all capable men had to take part in the war. Hence the

Table 1. Datasets providing Finnish historical occupations for AMMO.

Name	Data provider	Persons	Occupations
WW1 War Victims	National Archives	39 931	1391
WW2 Death Records	National Archives	94 700	2155
WW2 Prisoners of War	National Prisoners of War Project	4460	576

occupation distribution is also well balanced among social classes. Female occupations are not well represented, as the datasets contain only about 1340 female records.

3 Transforming Occupational Labels into an Ontology

The occupation literals were extracted from the datasets, and easily recognizable various spellings of worker occupations are grouped together programmatically, as well as almost identical occupational labels based on a Jaro-Winkler string similarity distance. This resulted in a flat vocabulary of 2053 distinct occupation groups that contained 2977 distinct labels. This data was used as the starting point for manual harmonization, structuring, and linking work using a spreadsheet that is transformed into RDF.

The occupation labels were manually studied and different spellings of the same occupation were harmonized, while quantifying whether the label actually refers to an occupation, or to a title or a social role. Occupations were linked to HISCO basic codes, and if applicable, to relation, status and product codes. Linking was also done to the Finnish Classification of Occupations 1980 (COO1980), which was also used as a controlled vocabulary to help in the harmonization and as a source of information on occupations' details. COO1980 is based on, and compatible with, the Nordic classification of occupations from 1963, which is based on ISCO-58, and so shares the same roots with HISCO. COO1980 is compatible with several 20th century national censuses.

The attribution of a HISCO code to each occupational label was carried out without further insight into the individual person records. In some cases, the occupational label is too general [4] to really understand what a person with that occupation label has been doing. For example, the occupation "hioja" (grinder or polisher) might correspond to several different categories of activities, related to, e.g., glass, cement, wood, metal, stone, or tools production. In these cases, the 99999 HISCO code was assigned (the title is too vague or ambiguous).

The HISCO relation codes were used for the abundant occupations with a family relation, such as "driver's child", "driver's widow" or "training school teacher's wife": such entries have a historical foundation in a specific cultural context and therefore should be registered as such. Women were often registered as wives with only the husband's activity mentioned. For example, "butcher's wife" is linked to "butcher" HISCO occupation code with the corresponding relation code (11). The HISCO status codes add information regarding the status of workers and, with the basic code, is important for determining the socioeconomic status: master artisan, apprentice, etc.

4 Discussion

This paper presented the foundational work in creating an ontology of Finnish historical occupations. AMMO will enable answering research questions relating to occupations and social stratification in historical datasets. For example, the

socioeconomic statuses of the two sides fighting in the Finnish Civil War can be compared. Combined with a historical place ontology, this can be further expanded to understand which social strata have joined either side in the war in different parts of the country.

The application of HISCO to a national reality presents complex tasks: creating a historical thesaurus of occupational names in the national language(s), possibly adapting the coding or creating additional coding to refine information. For example, for the Catalonian project [5] additional coding on institution was added to the administrative occupations, in order to distinguish between Colonial Administration, Army, Diplomacy, Local government administration, etc.

When assigning a HISCO code, it is important to find a balance between accuracy and generalization. In our case, HISCO coding is performed on secondary sources and therefore on occupational labels arranged into groups; the analysis is not performed on individual person records. In general, when an occupational label is too vague, further research on production activities in the place of residence and on familial occupational records should be done.

Our future work will study and analyze the aforementioned datasets with AMMO. AMMO is planned to be released in 2019, and will be used as a shared ontology for Finnish historical datasets in WarSampo and in an upcoming portal of the Finnish Civil War.

References

1. Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J., Mäkelä, E.: WarSampo Data Service and Semantic Portal for Publishing Linked Open Data about the Second World War History. In: The Semantic Web – Latest Advances and New Domains (ESWC 2016). pp. 758–773. Springer-Verlag (2016)
2. Koho, M., Gasbarra, L., Tuominen, J., Rantala, H., Jokipii, I., Hyvönen, E.: AMMO Ontology of Finnish Historical Occupations. In: Proceedings of the The First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH'19) (2019)
3. Mandemakers, K., Mourits, R.J., Muurling, S., Boter, C., van Dijk, I.K., Maas, I., de Putte, B.V., Zijdeman, R.L., Lambert, P., van Leeuwen, M.H., van Poppel, F., Miles, A.: HSN standardized, HISCO-coded and classified occupational titles, release 2018.01. IISG, Amsterdam (2018)
4. Matthijs, K., Peeters, H., Van den Troost, A., Van de Velde, I.: The coding of 19th century occupations from three different Belgian regions into ISCO68. KU Leuven, Departement Sociologie; Leuven (1998)
5. Pujadas-Mora, J.M., Marín, J.R., Villar, C.: Propuestas metodológicas para la aplicación de HISCO en el caso de Cataluña, siglos XV-XX. Revista de Demografía Histórica **32**, 181–219 (2014)
6. Statistics Finland: Classification of Occupations 1980. Statistics Finland (1981)
7. Van Leeuwen, M.H.D., Maas, I., Miles, A.: HISCO: Historical international standard classification of occupations. Leuven University Press (2002)
8. Zijdeman, R., Lambert, P.: Measuring social structure in the past: A comparison of historical class schemes and occupational stratification scales on dutch 19th and early 20th century data. Journal of Belgian History **40**(1-2), 111–141 (2010)