AMMO ontology of Finnish historical occupations

Mikko Koho[1] mikko.koho@aalto.fi
Lia Gasbarra[1] lia.gasbarra@aalto.fi
Jouni Tuominen[2,1] jouni.tuominen@helsinki.fi
Heikki Rantala[1] heikki.rantala@aalto.fi
Ilkka Jokipii[3,4] ilkka.jokipii@arkisto.fi
Eero Hyvönen[2,1] eero.hyvonen@helsinki.fi

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

Abstract

This paper introduces AMMO Ontology of Finnish Historical Occupations. AMMO is based on thousands of occupational labels extracted from three Finnish military historical datasets of the early 20th century: the first consists of the ca. 40 000 war-related death records around the time of the Finnish Civil War (1914–1922); the second consists of the ca. 95 000 death records of Finnish soldiers in the Second World War (1939–1945); the third contains the ca. 4500 records of Finnish prisoners of war in the Soviet Union during the WW2. Our goal from a Digital Humanities perspective is to use AMMO to study military history and these datasets based on the occupation and social status of the soldiers. AMMO will also be used as a component for faceted search and semantic recommendation in two semantic portals for Finnish military history. AMMO is aligned with the international historical occupation classification HISCO and with a modern Finnish occupational classification for international and national interoperability. The ontology is published as Linked Open Data in an ontology service.

1 Introduction

The measurement of historical social stratification has been a source of discussion in social history studies in the last decades [14]. Whether skills, capital, property, nobility, or occupational prestige is a suitable measure of social status has been debated: often researchers work with vague occupational information, as occupational labels are unclear, and there might not be enough context information to understand the reality of the people working in the occupation.

After extensive research on large historical datasets, the HISCO historical international standard classification of occupations [19] was published in 2002. It provides an international comparative classification system of history of work, particularly for occupational titles in the 19th and early 20th centuries. HISCO encodes not only occupation, but also information about prestige, property and family relations can be included. In general, the national classifications of occupations or census tables, differ in structure and detail within a country, and especially in international context. HISCO provides a tool for transnational comparative studies while also enabling the harmonization of occupations in censuses and datasets on a national scale.

In: Carlo Meghini, Antonella Poggi (eds.): Proceedings of ODOCH 2019 – Open Data and Ontologies for Cultural Heritage – Rome, Italy, 3 June 2019, published at http://ceur-ws.org

AMMO ontology will provide a harmonized view of Finnish historical occupations, which is linked to HISCO classification. The AMMO background and involved manual expert work has been discussed in a previous publication [2]. This paper builds upon the previous work to present the processes used to create the ontology, the ontology design rationale, and the ontology model. HISCO provides the hierarchical backbone of occupational groups in AMMO, as well as social stratification information through several measures like HISCLASS [18,11], a HISCO-based 12 level social classification system, and HISCAM [10,11], a social interaction distance measure. AMMO is also aligned with the Finnish Classification of occupations 1980 [17] (COO1980), a social stratification classification system in use in Finland.

AMMO ontology is based on occupational labels extracted from three Finnish military historical datasets of the early 20th century: the first consists of the ca. 40 000 war-related deaths around the time of the Finnish Civil War (1914–1922)[1]; the second consists of the ca. 95 000 death records of Finnish soldiers in the Winter War and Continuation War (1939–1945) [8]; the third contains the ca. 4500 records of the Finnish prisoners of war in the Soviet Union during the WW2. The two latter are part of the WarSampo[2] data service and semantic portal [4].

Motivation for AMMO comes from two separate usage scenarios. First is using the occupations in a user interface with a faceted search and the second is performing historical research on datasets consisting of data about people. Using the raw occupational labels does not enable the selection of person records based e.g. on the occupational field, social status, and various spellings of a single occupation. These issues can be solved by organizing the occupational labels into an ontology and linking to classifications with information on the social status of the occupation.

The benefits of an occupation ontology in the two scenarios can be summarized as follows:

- **User-interfaces.** User-interfaces employing faceted search [13] (e.g. semantic portals) benefit from organizing each facets' selection into a controlled vocabulary. This holds also for other user interface designs that list or show all of the values within a dataset to a user. Occupations are one of the key variables in many fields of history [19], and thus are one of the natural facets when exploring, studying and analyzing a dataset consisting of people. Using an ontology of occupations enables showing and using hierarchical facet options, and to group synonyms together into a single option and separate homonyms into separate options. Combined with information on the social stratification related to each occupation, we are able to create additional facets based on the social classes.
- **Historical research.** Digital Humanities researchers studying and researching history can use the ontology to get more understanding about the social stratification and occupational distribution within a dataset. Combined with ontology-based query expansion [22], the ontology enables the selection and comparative study of people and their information, based on arbitrary grouping resources, like the occupational field and social class. Also, these prosopographical groups [20] can be enriched with information like the average social class, and the most common occupational field. Many of the research questions of a collaborating historian revolve around social stratification, which is feasible to study only after linking the occupational labels to social stratification measures or classes. An example of the research questions we are trying to answer is ”what is the difference in the social stratification of the two sides fighting in the Finnish Civil War? Which social strata have joined either side in the war in different parts of the country?”

Our work is based on earlier studies about classifying occupations and social stratification. We strive to use pre-existing classifications as much as possible, so surveying the existing occupation classifications has been fruitful, and it sets the limits of the work, as manual expert work on vocabularies is time-consuming.

2 Existing Classifications of Historical Occupations

HISCO is based on a pre-existing international classification of occupations: ISCO-68, which in many countries has been adopted as a guideline for the creation of a national occupation classification scheme. In that case, the aligning of a pre-existing national classification scheme into HISCO is less problematic, since the structures are similar and entries are easily comparable. There is a Finnish version of the Nordic classification of occupations from 1963 [9], based on ISCO-58, which the ISCO-68 is based on. A newer Finnish classifications of occupations from 1980 [17], used e.g. in late 20th century census data, is in turn based on the aforementioned Nordic classification of occupations.

[1] http://www.ldf.fi/dataset/narc-sotasurmat1914-22

[2] http://www.ldf.fi/dataset/warsa

The HISCO encoding process in AMMO is carried out manually: occupational labels are linked to HISCO using the COO1980 as a reference, which is a consistent source of about 5100 specific occupational terms arranged hierarchically. The detailed occupational entries and description of the occupational groups in it helped to interpret and understand the numerous labels enough to enable the HISCO coding. Some occupations have required more specific attention, e.g. those with uncertain attribution such as *"keittäjä"*, cook, which presents many alternatives like *canteen cook*, *sugar cooker*, *sterilizing cook* or *pulp digester operator*, and distinct but hermetic occupational names, such as *"happomies"*, literally "acid man", which is a specific pulp industry worker.

Interesting sources of data for comparative studies are the national censuses of the early 20th century. These, however, group occupations under large, coarse categories, which are impossible to directly link to AMMO or HISCO, as the actual occupations are not known.

Other Finnish historical sources presenting listings of occupations are, for example, the yearly classifications of worker occupations for bread voucher distribution from 1940 [5] to 1943 [6] where working population was divided by occupation and the production sector. Population was ranked according to the grade of manual labour performed; harder labour corresponded to a higher class of bread, butter, and milk voucher. Specifically, the classification of 1943 [6] presents very detailed listings of occupations, accompanied by the corresponding value of the voucher. It is evident how the purpose of a classification influences its intrinsic structure and level of detail.

3 Creating an Ontology of Finnish Historical Occupations

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

The source datasets of AMMO are presented in Table 1, containing information of ca. 139 000 historical persons (soldiers), of which almost all are annotated with at least one occupational labels, summing up to thousands of different occupation titles. 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 label of a person is actually not an occupation but a social role, honorary title, degree or status, such as *student*, *nobleman*, *child*, *tenant* or *master of science*. Many children are labeled under their father's occupation, such as *driver's son*. Although occupations in HISCO are by definition solely activities that generate a remuneration, it is possible to also categorize many social roles or statuses through HISCO relation and status coding.

A common approach to creating an ontology model is to reuse existing non-ontological knowledge resources such as thesauri, classification schemes and lexicons, or ontological knowledge resources [21]. For AMMO, the existing Finnish classification of occupations 1980 [17] was used as both a thesaurus, and a classification scheme for the identification of both a specific occupation and a social status.

The main design rationale of the ontology model comes from the aforementioned two usage scenarios, i.e. the need to use occupational information in faceted search and historical research. We have striven to create the simplest possible model to provide results for these, that does not lose important information given in the occupational labels. The secondary goal is to provide a useful artifact for anyone studying or analyzing historical data containing people with Finnish language occupational labels.

One approach to achieving the needed HISCO-linking would be to annotate the HISCO occupations directly with the corresponding Finnish occupational labels found in our datasets. However, as the HISCO status and relationship variables are an important part of the HISCLASS coding [18], performing the coding on only HISCO occupation code would be erroneous for many occupational labels, as e.g. a pharmacy student would be considered having the same HISCLASS code as a pharmacist.

The AMMO ontology consists of individual SKOS concepts [12], each depicting one occupation with synonyms and alternative spellings gathered to the same concept as alternative labels. This enables to fully employ HISCLASS coding, and to keep the level of detail used in the occupational labels in the source datasets. The AMMO concepts are further separated into 5 classes depending on whether the occupational label refers to 1) an actual occupation, 2) a degree, 3) an honorary title, 4) a military rank, or 5) a social role.

The ontology model is presented in Figure 1 through two example resources, of which one is an occupation (pharmacist), and one is a social role related to the occupation (pharmacy student). RDF resources are depicted as ellipses, literals as rectangles and related datasets as clouds. The figure displays the linkage to the existing occupation classifications, and the related classification hierarchies. There are no direct relations between the AMMO occupation concepts, but the concepts (in green) are linked to other resources:

- HISCO (in blue), which contains the occupation hierarchy, relationship code, status code, HISCAM measure, and HISCLASS class,
- COO1980 (in red), which contains the occupation hierarchy, socioeconomic status class,
- KOKO ontology (in yellow).

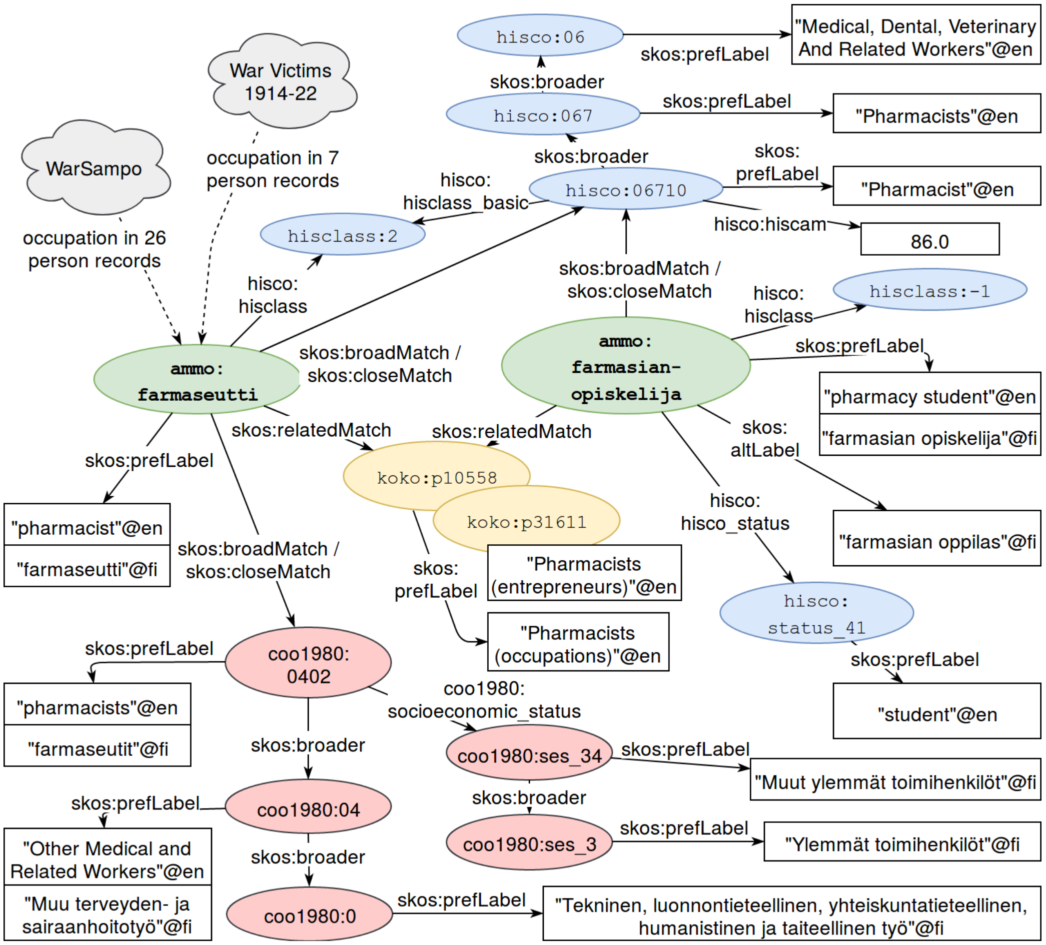

Fig. 1. The AMMO ontology model with two example AMMO resources.

In Figure 1, the namespace prefix *ammo:* refers to AMMO ontology namespace, *hisco:* refers to the RDF conversion of HISCO, *coo1980:* refers to the RDF conversion of COO1980, and *koko:* refers to the KOKO ontology. The property *hisco:hisclass* annotates the HISCLASS class code (1-12, or -1 for no occupation) of an AMMO occupation, whereas the base HISCLASS code of an HISCO occupation is given with the property *hisco:hisclass_basic*. There are two linked KOKO ontology concepts for both AMMO resources.

The overall process of creating the AMMO ontology is as follows:

1. Combining occupational labels from the datasets, and automatic grouping of easily identifiable synonyms,
2. The manual harmonization of the occupational labels and linking to external vocabularies,
3. Transforming the occupations into a SKOS vocabulary,
4. Validating and refining the ontology as needed,
5. Integrating HISCO and COO1980 classifications as linked SKOS vocabularies.

In step 1, the occupational labels are extracted from the datasets, and programmatically harmonized using a few simple rules to group occupational labels containing common interchangeable worker names *”työläinen”*, *"työmies"*, and *"työntekijä"*, which in most cases are used for identical meaning, and occupations with almost identical labels based on a Jaro-Winkler string similarity limit of 0.97. This results in a flat vocabulary of 2053 distinct occupations, containing a total of 2977 distinct occupational labels.

Step 2 begins with transforming the flat vocabulary into a spreadsheet, for an ontology developer to work on. The ontology developer re-engineers the ontology to account for synonymy, while manually linking the occupations to HISCO and COO1980 classifications and to the KOKO ontology[3].

Step 3 consists of RML [1] transformation of the spreadsheet into a SKOS [12] vocabulary. The ontology already contains URI references to the used classifications and the KOKO ontology, as well as annotations of preferred and alternative labels.

In step 4, the created ontology is validated and refined as needed. One key validation is to link the person records in the source datasets to AMMO, and inspect the results.

Step 5 consists of transforming the HISCO version 2018.01 [11] and the COO1980 main hierarchy into SKOS vocabularies, and enriching them with pre-existing English and Finnish labels. They are integrated into AMMO to provide hierarchical backbones, which might still reveal a need to refine the manual harmonization.

4 Discussion

This paper presented the foundations of the AMMO ontology, which will enable better utilization of Finnish historical datasets containing information about people. User interfaces can make use of either of the occupational hierarchies to provide a faceted search of people based on their occupation, in addition to enabling selection based on persons' social class, or e.g. the line of work (agriculture, metal workers, etc.). Historians can pursue research questions related to social stratification, line of work, and various occupational groups.

AMMO is an ontological representation of Finnish occupations, for the period ranging from 1914 to 1945, therefore having clear boundaries in space and time.

In order to link occupational names gathered from disparate sources into HISCO coding, the effort of interpretation and attentive adjustments are necessary, despite historically relevant occupational statistics and official classifications of occupations being readily available. The first half of the twentieth century saw substantial transformations in the Finnish society, especially in the agricultural sector: a long-lived vertical hierarchical system was shifting towards a more horizontal structure. The statuses of some agricultural occupations have changed dramatically while the occupation name has remained the same. This semantic drift [3] in the occupations causes the HISCO codings to be time-dependent, and HISCO coding based on occupations in the early 20th century might not be accurate in previous centuries. In addition to being a possible obstacle to some comparisons, the semantic drift provides an interesting topic to study in the future.

One interesting direction of research would be to compare the social stratification of people on different sides of the Finnish civil war with that of the social stratification on the national level. This would require at least to estimate a HISCLASS level to each coarse-grained occupational group.

Generally, Finland presents an ideal situation in population data availability and accuracy [7]. The first population census was completed already in 1749 under Swedish jurisdiction, after which they have been regularly redone [16]. Population registrations have been historically also registered in detail [15].

Currently work on AMMO is in step 3 of the process depicted in Section 3. Later, the AMMO ontology, along with the conversion pipeline will be published online for anyone to use.

References

1. Dimou, A., Sande, M.V., Slepicka, J., Szekely, P., Mannens, E., Knoblock, C., Walle, R.V.D.: Mapping hierarchical sources into RDF using the RML mapping language. Proceedings - 2014 IEEE International Conference on Semantic Computing, ICSC 2014 pp. 151–158 (2014). https://doi.org/10.1109/ICSC.2014.25
2. Gasbarra, L., Koho, M., Jokipii, I., Rantala, H., Hyvönen, E.: An ontology of finnish historical occupations. In: Proceedings of the 16th ESWC Conference (ESWC 2019), Posters & demonstrations. Springer (June 2019)
3. Gulla, J.A., Solskinnsbakk, G., Myrseth, P., Haderlein, V., Cerrato, O.: Semantic drift in ontologies. In: WEBIST (2). pp. 13–20 (2010)
4. 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)
5. Kansanhuoltoministeriö: Ohjeet leipä-, rasva- ja maitokorttien jakelusta. Kansanhuoltoministeriö, Helsinki (1940)

[3] http://finto.fi/koko/en/

6. Kansanhuoltoministeriö: Leipäkorttien jaossa noudatettava ammattiryhmittely. Kansanhuoltoministeriö, Helsinki (1943)
7. Kinnunen, M.: Luokiteltu sukupuoli. Vastapaino, Tampere (2001)
8. Koho, M., Hyvönen, E., Heino, E., Tuominen, J., Leskinen, P., Mäkelä, E.: Linked Death - Representing, Publishing, and Using Second World War Death Records as Linked Open Data. In: The Semantic Web: ESWC 2017 Satellite Events. pp. 369–383. Springer (2017)
9. Kulkulaitosten ja yleisten töiden ministeriön työvoima-asiain osasto: Pohjoismainen ammattiluokittelu, suomenkielinen laitos. Valtioneuvoston kirjapaino, Helsinki (1963)
10. Lambert, P., Zijdeman, R., Leeuwen, M.V., Maas, I., Prandy, K.: The construction of HISCAM: A stratification scale based on social interactions for historical comparative research. Historical Methods: A Journal of Quantitative and Interdisciplinary History **46**(2), 77–89 (2013)
11. 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)
12. Miles, A., Matthews, B., Wilson, M., Brickley, D.: SKOS core: simple knowledge organisation for the web. In: International Conference on Dublin Core and Metadata Applications. pp. 3–10 (2005)
13. Oren, E., Delbru, R., Decker, S.: Extending Faceted Navigation for RDF Data. In: The Semantic Web - ISWC 2006. pp. 559–572. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
14. Van de Putte, B., Buyst, E.: Occupational titles? hard to eat, easy to catch. Journal of Belgian History (JBH) **40**(1-2), 7–31 (2010)
15. Sköld, P.: The birth of population statistics in sweden. The History of the Family **9**(1), 5–21 (2004)
16. Statistics Finland: Historiallisen tilastotiedon opas. https://guides.stat.fi/historiallisentilastotiedonopas/vaestolaskennat, accessed: 2019-05-14
17. Statistics Finland: Classification of Occupations 1980. Käsikirjoja / Tilastokeskus, Statistics Finland, Helsinki (1981)
18. Van Leeuwen, M.H.D., Maas, I.: HISCLASS: A historical international social class scheme. Leuven University Press (2011)
19. Van Leeuwen, M.H.D., Maas, I., Miles, A.: HISCO: Historical international standard classification of occupations. Leuven University Press (2002)
20. Verboven, K., Carlier, M., Dumolyn, J.: A short manual to the art of prosopography. In: Prosopography approaches and applications. A handbook, pp. 35–70. Unit for Prosopographical Research (Linacre College) (2007)
21. Villazón-Terrazas, B.C., Suárez-Figueroa, M., Gómez-Pérez, A.: A pattern-based method for re-engineering non-ontological resources into ontologies. International Journal on Semantic Web and Information Systems (IJSWIS) **6**(4), 27–63 (2010)
22. Voorhees, E.M.: Query expansion using lexical-semantic relations. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 61–69. Springer-Verlag New York, Inc. (1994)