Studying Occupations and Social Measures of Perished Soldiers in WarSampo Linked Open Data

Mikko Koho[1], Eero Hyvönen[1,2]

[1]*Semantic Computing Research Group (SeCo), Aalto University, Espoo, Finland*
[2]*HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Helsinki, Finland*

Abstract

WarSampo Knowledge Graph contains data about Finland in the Second World War as Linked Open Data, including metadata of more than 100 000 people. A crucial part is the casualty register, containing detailed information about all Finnish soldiers who perished during the war, consisting of 94 676 person records. This register contains occupational labels, which have been manually harmonized into an ontology and linked to occupational classifications such as HISCO. This paper gives an overview of the harmonized occupation ontology and provides an outlook of how the ontology, the occupation classifications, and related social measures could be used for prosopographical study in the future to provide new insights into events of the war or of the surrounding society.

1. Introduction

WarSampo — Finnish Second World War (WW2) on the Semantic Web [1] collects, integrates, and harmonizes data about Finland in WW2 and publishes the resulting *Knowledge Graph (KG)* as *Linked Open Data (LOD)* [2]. The data is available via a public web portal[1] for searching, browsing, and analyzing the data through nine different perspectives, as well as an open SPARQL endpoint for direct data access. Since its opening in 2015, the WarSampo portal has been used by more than a million end users, corresponding to almost 20% of the population of Finland.

The core dataset of WarSampo is the casualty register [3] of the National Archives of Finland, containing detailed information about all Finnish soldiers who perished during the war, consisting of 94 676 person records. This data is enriched with interlinked datasets about, e.g., military units, war diaries, wartime photographs, historical places, historical maps, and war-time events. WarSampo also contains metadata of all known Finnish prisoners of war (POW) (4200 persons) [4] and notable individuals (5611) from other data sources, such as Wikipedia.

The WarSampo KG enables seeking new insights about WW2, the arguably most devastating catastrophe in human history. The ontology and data infrastructure of WarSampo can be further

[1]WarSampo Portal: http://sotasampo.fi/en/

extended with new data to enable digging even deeper into the societal research questions which interest many military history scholars today [5, 6]. Towards this end, considerable effort was put in harmonizing the occupational labels in the person registers [7] to enable prosopographical study of the persons based on their occupational groups and social measures. The devised occupation ontology AMMO [8] is aligned with and linked to the *HISCO* historical international standard classification of occupations [9]. HISCO provides the hierarchical backbone of occupational groups in AMMO, as well as two social measures [10]: *HISCLASS* [11] for social classes and *HISCAM* [11] for continuous scale social status. AMMO is also aligned with the Finnish *Classification of occupations 1980 (COO1980)* [12], a occupation classification system which has been in use in Finland.

With the enhanced possibilities for information retrieval and data grouping attained from the harmonization and reconciliation of occupations into rich ontologies, the possibilities for answering humanities-driven research questions are enhanced. Exploiting the new possibilities requires understanding about the data provenance, Semantic Web technologies, and computational data analysis, as well as domain knowledge of military historical research, thus making it an interesting case for interdisciplinary Digital Humanities research [13, 14].

This paper extends our earlier publications on WarSampo by giving a statistical overview of the harmonized occupation ontology and providing an outlook of how the ontology, the linked occupation classifications, and related social measures could be used for prosopographical study in the future to provide new insights into events of the war or of the surrounding society. The data harmonization processes of AMMO ontology have been presented in [7] and the ontology structure has been presented in [8]. A previous paper [15] has shown how various prosopographical phenomena can be highlighted and visualized. We extend this line of research by focusing on studying the occupations and social measures of the perished soldiers.

2. WarSampo Knowledge Graph and LOD Infrastructure

In WarSampo, Linked Data and the event-based *CIDOC Conceptual Reference Model (CRM)* [16] are used as a basis for harmonizing datasets about Finland in the Second World War into a unified KG [2]. Main entity types in the KG are persons, military units, death records, prisoner records, events, places, photographs, war diaries, articles, and occupations. The death and prisoner records were created from the metadata records of the casualty and POW databases of the National Archives, respectively, and were aligned with the WarSampo KG person entities.

The death record data is of great importance in studying Finland in WW2, as it contains detailed information about all 94 676 perished soldiers in the Finnish fronts[2]. The POW register contains data about all 4200 Finnish POWs. In addition, WarSampo contains information of over 5600 notable persons who survived the war, aggregated from additional data sources.

The occupations of the person registers have been harmonized into an occupation ontology [8], which is linked to HISCO historical international standard classification of occupations [9]. HISCO provides the hierarchical backbone of occupational groups in AMMO, as well as social stratification information through several measures like HISCLASS [17, 11], a HISCO-based 12 level social classification system, and HISCAM [18, 11], a social interaction distance measure.

---

[2]As far as they are known, excluding some exceptions for data privacy reasons, such as deaths by suicide.

AMMO is also aligned with the Finnish Classification of occupations 1980 [12] (COO1980), a social stratification classification system in use in Finland.

There are also person related documents that are linked to the person instances or their military units, including a large collection of some 164 000 wartime photographs, tens of thousands hand-written digitized war diaries, and thousands war veteran magazine articles. These provide further contextual information for people studying, for example, the war paths of their relatives. The latest version of the WarSampo KG is always available at the Linked Data Finland LDF.fi platform[3] with its SPARQL endpoint and other services [19] and at Zenodo [20].

Using Semantic Web technologies and CIDOC CRM help to create a sustainable and collaborative infrastructure for pursuing historical research [21]. Anyone can link their data to the WarSampo entities, and enrich their data from WarSampo. For example, the domain ontology of people provides a point of access to all of the information about each person contained in WarSampo, making it possible to for anyone to use this information by linking to the person.

Many of the domain ontologies of WarSampo, e.g., occupations, military ranks, and war-time municipalities, are used to provide facet values in the many faceted search perspectives of the WarSampo portal. These can be re-used to enable faceted search with other datasets. The faceted search user interfaces of the WarSampo Portal provide an easy way for anyone interested in military history to study, explore, and analyze the integrated knowledge graph [4]. For example, one can do simple prosopographical data analysis of the person registers.

The Casualties perspective of the portal makes use of the harmonized occupational titles of AMMO, but does not currently use or display the social stratification measures. It would be possible, however, to add new facets for occupational groups or associated social classes.

In addition to studying the occupations through the Casualties perspective and Prisoners of War perspective, the portal contains a landing page for each person, which reconstructs the person's biography based on all of the information relating to a person from different sources [4], including his or her occupations. An example of a person's landing page is shown in Figure 1.

The WarSampo infrastructure has recently been employed in the WarMemoirSampo system[4] to provide named entities and contextual information to the things being discussed in war veteran interview videos, such as places, organizations, persons, military units, and events [22].

3. Occupations and Social Measures in AMMO Ontology

AMMO[5] contains currently 2258 occupations. These all have HISCO codes attached and 2152 occupations have an COO1980 code, too. The ontology is published on an open SPARQL endpoint[6] and on the ONKI ontology service[7].

The occupational labels from the source data are grouped to AMMO occupations when there seems to be no semantic difference between several labels. Multiple occupations can also use the same HISCO codes, which means that they would be identical in HISCO terms, unless

---

Figure 1: The landing page of a person, showing biographical information from various sources, including his occupations.

they make use of the HISCO relation, status, or product codes, which are used to describe the occupation in HISCO.

3.1. HISCO

HISCO contains a hierarchy of 1675 distinct occupational categories, with the purpose to be a classification system that enables comparisons to be made internationally and historically [10]. HISCAM is a continuous historical status scale that is linked to the occupational categories, as is the HISCLASS historical international social class scheme. Several versions of HISCAM exist based on different geographical and temporal datasets. HISCLASS contains 12 social classes, but also simplified versions of 7 and 5 levels are used.

The 12 most commonly used HISCO codes are shown in Table 1, with the total numbers of relation, status, and product codes used for the occupations. The most common occupations seen reflect the variety of occupational labels found in the source person registers relating to the corresponding HISCO occupation codes as well as to the decisions taken in interpreting the labels when creating the occupation ontology and deciding whether several occupational labels

can be considered the same occupation in the ontology or whether they are separate occupations but still refer to the same HISCO code. So the HISCO occupation code counts are somehow related to the prevalence of corresponding occupations in the source person registers. However, it is worth noting that for administrative and managerial professions, the exact field of work is often explicitly given in the occupational label, e.g., "bookbindery manager", "bookprinter manager", "car paint shop owner", and "rubber-repairshop's chief" .

Table 1
The 12 most common HISCO occupation codes in AMMO with the number of occupations referring to the code, along with the numbers of HISCO relation, status, and product codes used for the corresponding occupations

HISCO	HISCO label	Occupations	Relations	Statuses	Products
6-11.10	General Farmer	58	26	30	0
9-99.99	*Title is too vague*	46	13	4	0
-1	*Not an occupational title*	38	28	8	0
9-99.30	Factory Worker	36	10	0	0
6-21.05	Farm Worker, General	26	5	6	0
8-31.10	Blacksmith, General	24	9	1	0
6-32.20	Forest Supervisor	23	12	2	0
6-11.15	Small Subsistence Farmer (Husbandman)	23	9	13	0
3-93.10	Office Clerk, General	20	6	1	1
4-51.30	Retail Trade Salesperson	19	5	0	5
2-11.10	General Manager	19	6	2	0
8-49.10	Machinery Mechanic, General	19	3	1	1
8-55.10	Electrician, General	19	10	0	0

3.2. Classification of Occupations 1980

The Classification of Occupations (1980) contains 5100 specific Finnish occupational terms arranged hierarchically. It was published in 1980 and does not have a historical dimension but attempts to represent the contemporary occupations in Finland at the time. This classification has roots in the contemporary ISCO classification of the late 20th century, as does HISCO. COO1980 is compatible with several 20th century national censuses. It contains a 5-level social status classification for each occupation, similar to HISCLASS-5.

A look at the most common occupational groups of COO1980 in AMMO are shown in Table 2, which shows the middle-level occupational groups in the three level hierarchy used in COO1980 with the corresponding occupation counts. By surveying the occupations a bit closer, it is evident that there are some errors made in the manual harmonization process. For example, the code 30 "Managerial work in agriculture, forestry and fishing" seems to be the second most common group due to the fact that several occupations of the group with code 31 "Agricultural and horticultural work, animal husbandry" have accidentally ended up in the group 30, such as "livestock caretaker" and "raindeer herder". The code 91 "Occupation not specified" contains occupational labels that are too vague to fit into any other group, e.g., "workman" without any

specialization.

COO1980	Group label	Occupations
65	Iron and metalware work	162
30	Managerial work in agriculture, forestry and fishing	128
91	*Occupation not specified*	87
67	Wood work	86
11	Administration of private enterprises and organizations	83
72	Food and beverage work	71
31	Agricultural and horticultural work, animal husbandry	69
80	Public safety and protection work	68
63	Smelting, metallurgical and foundry work	65
10	Public administration	59
01	Supervision and executive work in the technical field	59

Due to unforeseen issues with the HISCO source data used in AMMO [8], a large amount of HISCO resources missed metadata in AMMO. For the study of this paper, we have enriched the data with the missing HISCO metadata and additional HISCAM measures provided by the DataLegend project on a SPARQL endpoint [23]. The data in that endpoint does not seem to be perfect either, as it is missing some of the higher level occupational groups, such as the codes "62" and "84".

4. Occupations in the Casualties Register

The casualty register contain $86\,069$ person records with occupational labels (of $94\,676$ person records in total). Most records have one occupation, but several records have two or more. The most common occupations in the register are shown Table 3. The HISCLASS column shows the HISCLASS-12 values of the occupations, with level 1 being the elite of the society and 12 being the lowest social class. The -1 (no HISCLASS code) values highlight the issues with mapping occupational labels to pre-existing classifications that rely on knowing the field of work on some level. The two generic worker occupations "työmies" and "sekatyömies" and the general student occupation can not be mapped to any HISCO occupation category. However, it might be possible to map them to 5 or 7 level HISCLASS classification. The very generic worker or manual laborer titles are a known common issue also with census data [10].

Figure 2 shows how the HISCAM and HISCLASS-12 values of the occupations in the casualty register correlate. It is easy to see a correlation with the HISCLASS class, but that correlation is not linear. Many occupations with -1 HISCLASS value have a HISCAM value.

5. Discussion

This paper has given an overview of the harmonized occupation ontology and an outlook of how the ontology, the occupation classifications, and related social measures could be used to

Table 3
The 10 most common occupations in the Casualties register.

URI	Label (fin)	Label (eng)	HISCLASS	Persons
http://ldf.fi/ammo/maanviljelija	maanviljelijä	farmer	8	22 111
http://ldf.fi/ammo/tyomies	työmies	worker	−1	15 319
http://ldf.fi/ammo/sekatyomies	sekatyömies	worker	−1	4761
http://ldf.fi/ammo/maanviljelijan-poika	maanviljelijän poika	farmer's son	8	4602
http://ldf.fi/ammo/maatyomies	maatyömies	farm worker	12	4010
http://ldf.fi/ammo/metsatyomies	metsätyömies	forest worker	12	2181
http://ldf.fi/ammo/autonkuljettaja	autonkuljettaja	driver	9	1523
http://ldf.fi/ammo/talollisen-poika	talollisen poika	farmer's son	8	1447
http://ldf.fi/ammo/opiskelija	opiskelija	student	−1	1364
http://ldf.fi/ammo/kirvesmies	kirvesmies	carpenter	7	1021
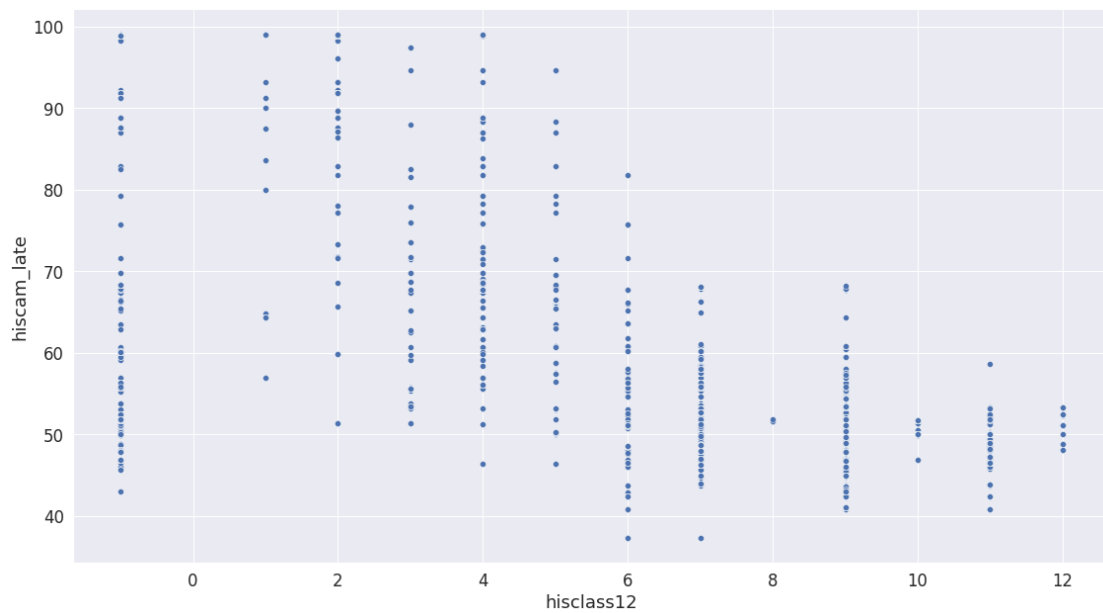
Figure 2: The HISCAM and HISCLASS values of occupations in the Casualties register.

study the WarSampo person registers. We are in the process of updating the AMMO ontology and the issues found in this study will be addressed in the next version. Future work will look at how the ontology can be used for prosopographical study using the WarSampo person registers to provide new insights into events of the war or of the surrounding society. HISCO is the most important and widely used standard for historical occupations [10] and the related social measures HISCAM and HISCLASS can be used to study social status and social classes.

In addition to WarSampo, AMMO is used in several other projects to provide occupational groups and information about social stratification, making it an important resource for Digital

Humanities research. It is planned to be taken into use in upcoming projects dealing with historical information about Finland. As the current focus is only on early 20th century, one topic of future research is handling occupations changing in time, for example through different interlinked versions of occupations for different time spans.

Wikidata[8] and Finnish KANTO – National Agent Data[9] are becoming relevant LOD sources for occupational labels for historical persons in addition to historical person registers and census data. One aspect of future work is to study and work on the compatibility between AMMO and these sources.

Acknowledgments

References

[1] E. Hyvönen, E. Heino, P. Leskinen, E. Ikkala, M. Koho, M. Tamper, J. Tuominen, E. Mäkelä, WarSampo data service and semantic portal for publishing linked open data about the second world war history, in: H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S. P. Ponzetto, C. Lange (Eds.), The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016, volume 9678 of *Lecture Notes in Computer Science*, Springer, Cham, 2016, pp. 758–773. doi:10.1007/978-3-319-34129-3_46.

[2] M. Koho, E. Ikkala, P. Leskinen, M. Tamper, J. Tuominen, E. Hyvönen, WarSampo knowledge graph: Finland in the second world war as linked open data, Semantic Web 12 (2021) 265–278. URL: https://doi.org/10.3233/SW-200392. doi:10.3233/SW-200392.

[3] M. Koho, E. Hyvönen, E. Heino, J. Tuominen, P. Leskinen, E. Mäkelä, Linked death — representing, publishing, and using Second World War death records as linked open data, in: E. Blomqvist, K. Hose, H. Paulheim, A. Ławrynowicz, F. Ciravegna, O. Hartig (Eds.), The Semantic Web: ESWC 2017 Satellite Events, volume 10577 of *Lecture Notes in Computer Science*, Springer, Cham, 2017, pp. 369–383. doi:10.1007/978-3-319-70407-4_45.

[4] M. Koho, E. Ikkala, E. Hyvönen, Reassembling the lives of Finnish prisoners of the Second World War on the Semantic Web, in: Proceedings of the Third Conference on Biographical Data in a Digital World (BD 2019), volume 3152, CEUR Workshop Proceedings, 2022, pp. 31–39. URL: http://ceur-ws.org/Vol-3152/BD2019_paper_5.pdf.

[5] T. D. Biddle, R. M. Citino, The role of military history in the contemporary academy, Foreign Policy Research Institute Footnotes (2015) 1–6. URL: https://www.fpri.org/docs/society_for_mil_hist_whit_paper.pdf.

[6] J. Black, Rethinking Military History, Routledge, 2004.

[7] L. Gasbarra, M. Koho, I. Jokipii, H. Rantala, E. Hyvönen, An ontology of Finnish historical occupations, in: P. Hitzler, S. Kirrane, O. Hartig, V. de Boer, M.-E. Vidal, M. Maleshkova, S. Schlobach, K. Hammar, N. Lasierra, S. Stadtmüller, K. Hose, R. Verborgh (Eds.), The Semantic Web: ESWC 2019 Satellite Events, Lecture Notes in Computer Science, Springer,

---

[8]Wikidata: https://www.wikidata.org/

[9]Kanto – National Agent Data: https://www.kiwi.fi/display/Toimijakuvailupalvelu/About+Kanto+in+English

Cham, 2019, pp. 64–68. URL: https://link.springer.com/chapter/10.1007/978-3-030-32327-1_13.

[8] M. Koho, L. Gasbarra, J. Tuominen, H. Rantala, I. Jokipii, E. Hyvönen, AMMO ontology of Finnish historical occupations, in: A. Poggi (Ed.), Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage, volume 2375 of *CEUR Workshop Proceedings*, 2019, pp. 91–96.

[9] M. H. D. Van Leeuwen, I. Maas, A. Miles, HISCO: Historical International Standard Classification of Occupations, Leuven University Press, 2002.

[10] M. H. D. Van Leeuwen, Studying long-term changes in the economy and society using the HISCO family of occupational measures, in: Oxford Research Encyclopedia of Economics and Finance, Oxford University Press, 2020. doi:10.1093/acrefore/9780190625979.013.541.

[11] K. Mandemakers, R. J. Mourits, S. Muurling, C. Boter, I. K. van Dijk, I. Maas, B. V. de Putte, R. L. Zijdeman, P. Lambert, M. H. D. Van Leeuwen, F. van Poppel, A. Miles, HSN standardized, HISCO-coded and classified occupational titles, release 2018.01, IISG, Amsterdam, 2018.

[12] Statistics Finland, Classification of Occupations 1980, Käsikirjoja / Tilastokeskus, Statistics Finland, Helsinki, 1981.

[13] S. Graham, I. Milligan, S. Weingart, Exploring big historical data. The historian's macroscope, Imperial College Press, London, UK, 2015. doi:10.1142/p981.

[14] A. Burdick, J. Drucker, P. Lunenfeld, T. Presner, J. Schnapp, Digital Humanities, The MIT Press, 2012.

[15] M. Koho, H. Rantala, E. Hyvönen, Digital humanities and military history: Analyzing casualties of the warsampo knowledge graph, in: K. Berglund, M. L. Mela, I. Zwart (Eds.), DHNB 2022 The 6th Digital Humanities in Nordic and Baltic Countries Conference, volume 3232, CEUR Workshop Proceedings, 2022. URL: http://ceur-ws.org/Vol-3232/paper29.pdf.

[16] M. Doerr, The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata, AI Magazine 24 (2003) 75–92. doi:10.1609/aimag.v24i3.1720.

[17] M. H. D. Van Leeuwen, I. Maas, HISCLASS: A historical international social class scheme, Leuven University Press, 2011.

[18] P. Lambert, R. Zijdeman, M. H. D. Van Leeuwen, I. Maas, K. Prandy, 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 (2013) 77–89.

[19] E. Hyvönen, J. Tuominen, M. Alonen, E. Mäkelä, Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets, in: ESWC 2014: The Semantic Web: ESWC 2014 Satellite Events, 2014, pp. 226–230. doi:10.1007/978-3-319-11955-7_24.

[20] M. Koho, E. Heino, P. Leskinen, E. Ikkala, M. Tamper, K. Apajalahti, J. Tuominen, E. Mäkelä, E. Hyvönen, WarSampo knowledge graph [data set], 2019. URL: https://doi.org/10.5281/zenodo.3611322. doi:10.5281/zenodo.3611322.

[21] D. Oldman, M. Doeer, S. Gradmann, Zen and the art of Linked Data: new strategies for a Semantic Web of humanist knowledge, in: S. Schreibman, R. Siemens, J. Unsworth

(Eds.), A New Companion to Digital Humanities, John Wiley and Sons, 2016, pp. 251–273. doi:`10.1002/9781118680605.ch18`.

[22] R. Leal, H. Rantala, M. Koho, E. Ikkala, M. Merenmies, E. Hyvönen, WarMemoirSampo: A semantic portal for war veteran interview videos, in: DHNB 2022. The 6th Digital Humanities in Nordic and Baltic Countries Conference, volume 3232, CEUR Workshop Proceedings, 2022. URL: http://ceur-ws.org/Vol-3232/paper30.pdf.

[23] R. Hoekstra, A. Meroño-Peñuela, A. Rijpma, R. Zijdeman, A. Ashkpour, K. Dentler, I. Zandhuis, L. Rietveld, The datalegend ecosystem for historical statistics, Journal of Web Semantics 50 (2018) 49–61.