Digital Humanities and Military History: Analyzing Casualties of the WarSampo Knowledge Graph

Mikko Koho[1,2], Heikki Rantala[1,2] and 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

This paper shows how various prosopographical phenomena can be highlighted and visualized in the WarSampo Knowledge Graph that contains rich data about Finland in the Second World War as Linked Open Data, including detailed metadata of more than 100 000 people. WarSampo Portal contains tools for simple prosopographical data analysis of the person registers, and accessing the SPARQL endpoint directly opens up further possibilities in using the ontology infrastructure for enhanced information retrieval and pursuing digital humanities studies. This paper overviews of these possibilities of WarSampo, and presents examples of how it, and by extension Linked Data more generally, can be used to created data analyses to support historical research.

Keywords

Military History, Linked Data, Digital Humanities, Data Analysis, Data Visualization, Prosopography

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)*, which is also part of the international *Linked Open Data Cloud*[1]. The data is available via an open SPARQL endpoint and a public web portal[2] for searching, browsing, and analyzing the data through nine different perspectives.

The core dataset of WarSampo [2] is the casualty register [3] of the National Archives of Finland, containing detailed information about all Finnish soldiers who were killed in action, consisting of 94 676 person records. This data is enriched with interlinked datasets of, 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 (Wikipedia, etc.) [5].

WarSampo Portal contains tools for simple prosopographical data analysis of the person registers. Accessing the SPARQL endpoint directly enables further analysis and retrieval of data

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022

mikko.koho@aalto.fi (M. Koho); heikki.rantala@aalto.fi (H. Rantala); eero.hyvonen@aalto.fi (E. Hyvönen)
https://seco.cs.aalto.fi/u/mkoho/ (M. Koho)
0000-0002-7373-9338 (M. Koho); 0000-0002-4716-6564 (H. Rantala); 0000-0001-7116-9338 (E. Hyvönen)

[1]Linked Open Data Cloud: https://lod-cloud.net/
[2]WarSampo Portal: http://sotasampo.fi/en/

for external tools. Just a SPARQL query and result set visualizations already enable answering complex questions about history. For example, one can study how the ratio of officers in perished soldiers differed geographically. The portal provides nine interactive "perspectives" on the data: (War) Events, Persons, Army Units, Places, Magazine Articles, Casualties, Photographs, War Cemeteries, and Prisoners of War. 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.

Semantic Web technologies provide viable solutions for combining heterogeneous isolated historical datasets [6]. With the enhanced possibilities for information retrieval and data grouping attained from the harmonization and reconciliation of metadata values with rich ontologies, the possibilities for answering humanities-driven research questions are greatly increased. 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 [7, 8].

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 extended with new data to enable digging even deeper into the societal research questions which interest many military history scholars today [9, 10].

This paper extends our earlier publications on WarSampo by showing how the LOD service and the portal can be used in Digital Humanities data analyses. This paper shows how various prosopographical phenomena can be highlighted and visualized. For example, we show 1) how the data can be used to visualize geographical variance in the ratio between enlisted soldiers and officers, 2) how the perished soldiers' places of domicile correlate with their mortality, social class, and place of burial, and 3) how one's occupation affected the likelihood of surviving in POW camps.

2. WarSampo Knowledge Graph

In WarSampo, Linked Data and the event-based *CIDOC Conceptual Reference Model (CRM)* [11] 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 WarSampo KG person entities [12].

The death record data is of great importance in studying Finland in WW2, as it contains detailed information about all 94 700 perished soldiers in the Finnish fronts. The POW register contains data about all 4200 Finnish POWs. In addition, WarSampo contains information of over 5600 notable persons who survided the war, aggretaged from additional data sources.

The occupations of the person registers have been harmonized into an occupation ontology [13], which is linked to the international HISCO classification and its related occupational measures [14], HISCLASS and HISCAM.

There are also person related documents that are linked to the person instances or their military units, including a large collection of wartime photographs, hand-written digitized

war diaries, and 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 LDF.fi platform[3]. All versions are available from Zenodo, and the current 2.1.0 version [15] is used for the analysis examples in this paper.

2.1. Linked Open Data Infrastructure for WW2

Using Semantic Web technologies and CIDOC CRM help to create a sustainable and collaborative infrastructure for pursuing historical research [16]. 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., military ranks and war-time municipalities, are used to provide facet values in the many faceted search perspectives of WarSampo Portal. These can be re-used to enable faceted search with other datasets.

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

2.2. WarSampo Portal

A number of data analytical visualizations can be performed on the WarSampo portal [17, 4]. 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 datasets. A user can do his/her own analysis of the data, which would generally often be impossible, as detailed historical data tends to be siloed in the repositories of different organizations and researchers.

Figure 1 presents a screenshot of the soldier life paths as a Sankey diagram from the Casualties perspective [17]. It shows the life paths of 40 soldiers from where the soldiers were born (on the left), where they lived, where they died, and where they are buried. In this case the facets have been used to filter the casualties to only show persons buried in the war cemetery of the town Inari in Ivalo, Lapland.

Cemeteries, local histories, and family histories interest the wide public to study the provided historical information to find information related to their own lives. In addition academic researchers and military history enthusiasts study the data through the portal.

3. Using the Knowledge Graph for Prosopographical Studies

To go beyond the possibilities of the semantic portal, one can use the LOD directly from the underlying SPARQL endpoint hosting the KG. LOD makes it potentially much easier to share research data in a way that benefits all. The main drawback is the new technical skills that need to be acquired. This section shows some examples of how visualizations can be created for prosopographical analyses using the WarSampo KG.

[3]WarSampo Knowledge Graph: https://www.ldf.fi/dataset/warsa

[4]Portal online: https://sotamuistot.arkisto.fi; project home: https://seco.cs.aalto.fi/projects/war-memoirs/en/

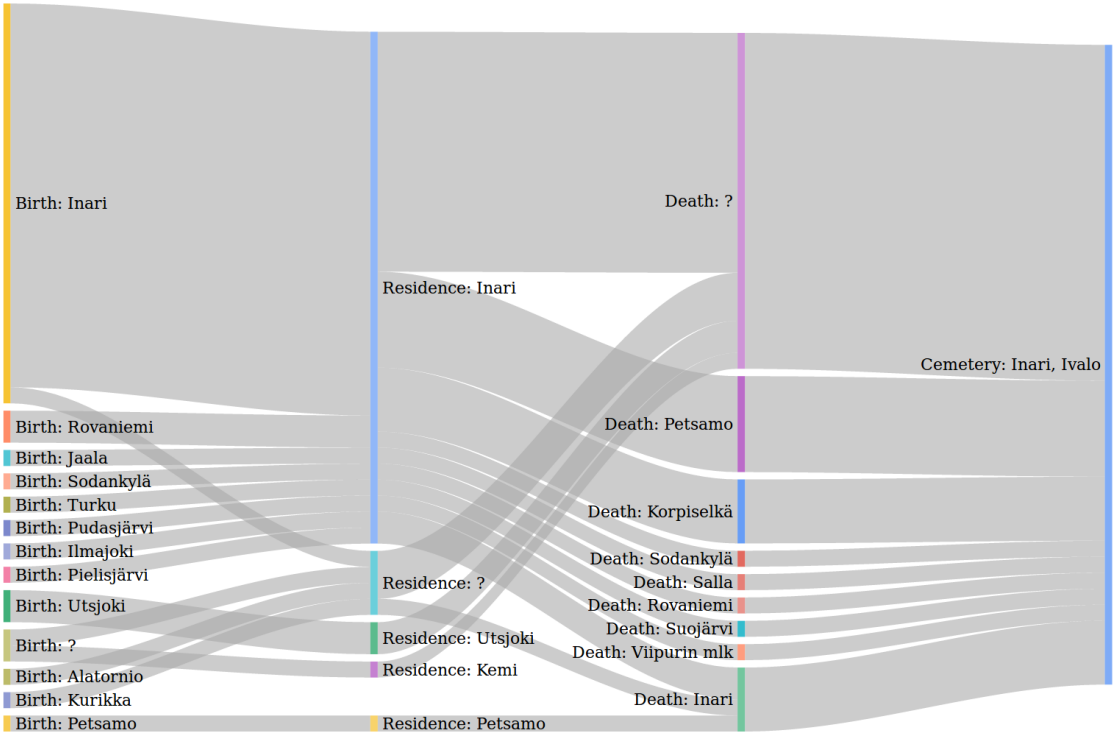

Figure 1: Life paths of the 40 soldiers in the war cemetery of Ivalo as shown in the Casualties perspective of WarSampo Portal. From left to right: Place of birth; place of residence; place of death; cemetery.

3.1. Daily Death Rates of Farmers During the Winter War

SPARQL query language is a major tool when using LOD. Below is an example of a relatively simple query[5] that will count the number of deaths among farmers, fishermen and others, based on the HISCLASS classification of their occupations. At the start of the query there is a number of PREFIX definitions that are not strictly necessary, but make the query easier to read and write. For example, the URI of the death record class *http://ldf.fi/schema/warsa/DeathRecord* can be written as *warsa:DeathRecord* after defining the prefix *warsa*. The SELECT line indicates the variables (identified by '?' character before variable name) that we want to extract from the data, in this case date of death and the number of deaths (per day). The section after WHERE include triple patterns that limit the query to the group that we are interested in, in this case people with HISCLASS 5-class scheme class 3 (farmers) who died during the Winter War. For example, the line '`?record a warsa:DeathRecord .`' is a triple pattern; it means that we want the variable '`?record`' to represent all the resources in the KG that are of type *http://ldf.fi/schema/warsa/DeathRecord*. We then limit the set of death records based on occupation and death date. At the end of the query we sort the results based on the dates in ascending order. These results that consist of dates and numbers can be visualized, e.g., as a line chart shown in Figure 2 created with the YASGUI editor tool [18].

[5]Yasgui can be used to share SPARQL queries; the query is here: https://api.triplydb.com/s/DCbeABJmE

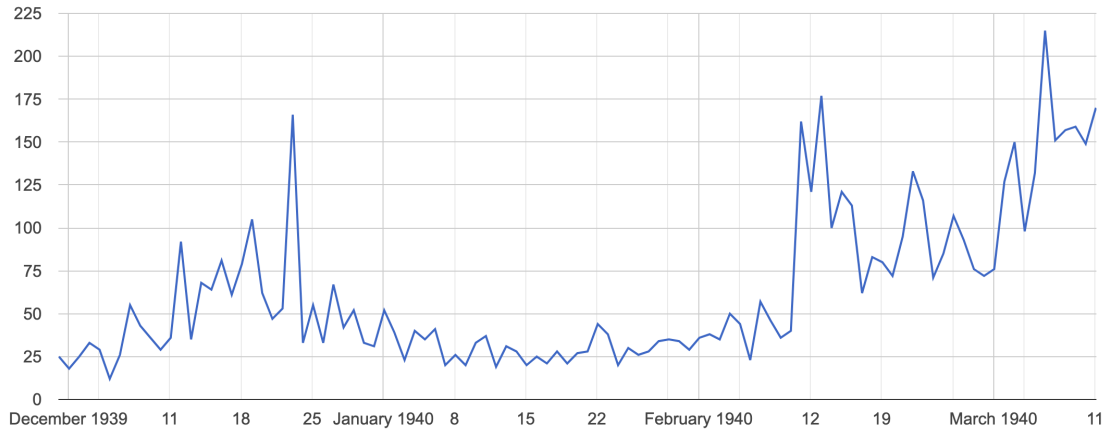

Figure 2: A line chart visualization showing the daily number of deaths among farmers during the Winter War.

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX warsa: <http://ldf.fi/schema/warsa/>
PREFIX casualties: <http://ldf.fi/schema/warsa/casualties/>
PREFIX bioc: <http://ldf.fi/schema/bioc/>
PREFIX ammo: <http://ldf.fi/schema/ammo/>

SELECT ?date_of_death (COUNT(?record) AS ?number_of_deaths) WHERE {
  ?record a warsa:DeathRecord .
  ?record bioc:has_occupation ?occupation .
  ?occupation ammo:hisclass5 <http://ldf.fi/ammo/hisco/hisclass5/3> .

  ?record warsa:date_of_death ?date_of_death .
  FILTER (?date_of_death >= "1939-11-30"^^xsd:date)
  FILTER (?date_of_death  <= "1940-03-11"^^xsd:date)
}
GROUP BY ?date_of_death
ORDER BY ASC(?date_of_death)
```

3.2. Ratio of Officers in Perished Soldiers

Often even somewhat complex visualizations can be created quickly with essentially only a SPARQL query. An example of this is can be seen in Figure 3. The bar chart visualization shows the proportion of perished officers to all perished soldiers in the data for each Finnish province based on the municipality of residence of each victim. It is easy to see that the Uusimaa Province ("Uudenmaan lääni") , where the Finnish capital Helsinki is located, has considerably higher ratio of officers than the other provinces.

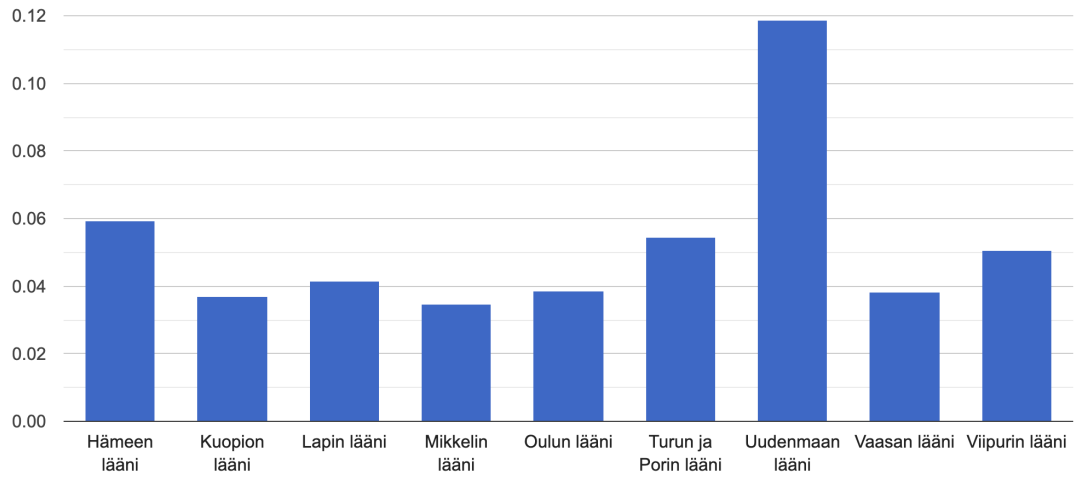

Figure 3: A bar chart visualization showing the ratio of officers in the Finnish perished soldiers, grouped by provinces of Finland based on the soldiers' municipality of domicile.

The visualization was created using the visualization tools of Yasgui and a short SPARQL query[6] of only a few lines. The places and ranks have simple hierarchical ontologies that are used to determine the province of residence and if the victim is an officer or not. This visualization uses only Finnish labels for the provinces as retrieved from the KG. However, LOD makes it easy to have names of entities in multiple languages using language tags.

3.3. Survival in Prisoner-of-War Camps

The register of the Finnish POWs is much smaller than the casualty register, but comparative studies with it can still provide interesting insights. Figure 4 shows the 7-class HISCLASS distribution of the prisoners in two groups: 1) those who survived the POW camps (blue), and 2) those who didn't survive (red), visualized with the Yasgui tool[7]. The large group "No HISCLASS code" corresponds mostly to non-specialized workers who worked in various tasks, often seasonally ("työmies" and "sekatyömies" in Finnish). There are 2726 prisoners with known occupation who survived and correspondingly there are 1393 perished prisoners.

One can see that there are differences in the distribution between the two groups in Figure 4. One can study this further by inspecting the group with the largest difference (omitting the "No HISCLASS code"): "Foremen and medium skilled workers". A visualization[8] of the 10 most common occupations in this group is shown in 5. This suggests that skilled workers had better chances of survival than others. As suggested by a collaborating historian, some skilled workers

[6]You can see and test this query here: https://api.triplydb.com/s/uKu3KbsQo
[7]The 7-class HISCLASS distribution of POWs: https://api.triplydb.com/s/DSwhf3w45
[8]Occupations of the group "Foremen and medium skilled workers": https://api.triplydb.com/s/arnMtmYEq

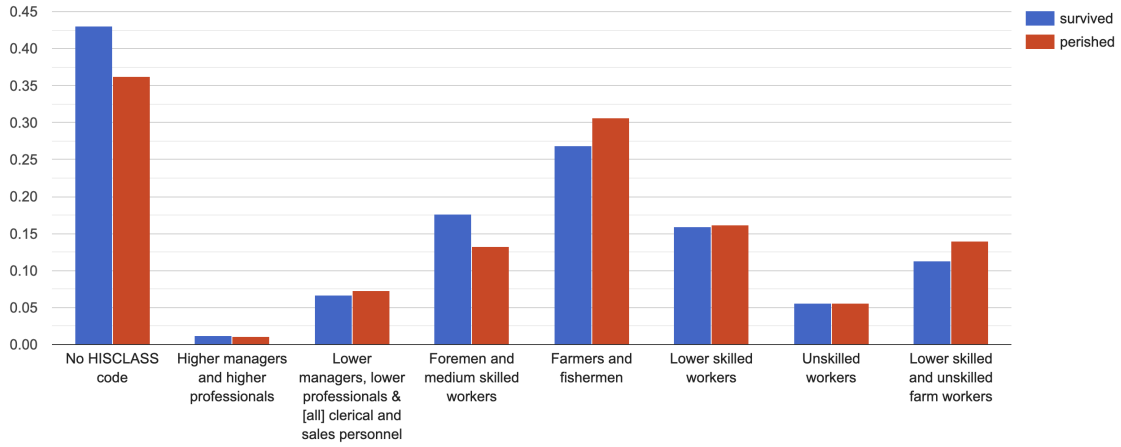

Figure 4: The distribution of 7-class HISCLASS classes of the POWs grouped by whether they survived in the POW camps or not.

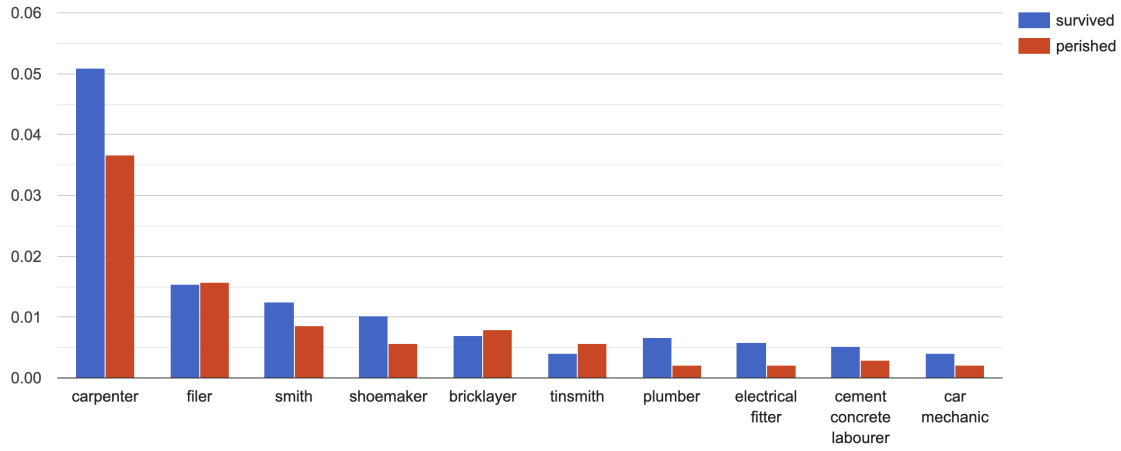

Figure 5: The POW occupation distribution in the HISCLASS group "Foremen and medium skilled workers" with 10 most common occupations shown, grouped by whether the soldiers survived the POW camps or not.

have been valued during their imprisonment as they have been useful in practical tasks, such as carpentry, and therefore might have received better treatment.

4. Discussion

In this paper we have shown how LOD can facilitate research in military history, as exemplified with the WarSampo KG. Metadata that is harmonized and reconciled into ontologies, like in the ontology infrastructure of WarSampo, provide enhanced possibilities for information retrieval that, combined with data analysis, can be fruitful for comparative prosopographical studies.

SPARQL is a potent tool for such studies once a suitable Knowledge Graph is available, but it is a tool that takes some effort to master for a person without computational skills.

To go beyond the portrayed examples, it would be crucial to pursue interdisciplinary Digital Humanities collaboration between Semantic Web researchers and humanist scholars. The former would provide the technical skill set needed for the studies, and the latter would provide the historical understanding of what would be interesting to study and how to interpret the produced results by setting them in context.

LOD enables new kinds of Digital Humanities studies by making it possible to easily do analysis that would be laborious, and often not feasible, to do otherwise. To use full potential of the possibilities required developing an ontology infrastructure to which the data is linked. LOD is most useful when integrating a number of heterogeneous, naturally interlinked datasets, providing a solid foundation for building such data infrastructures, that can be accessed and used by others.

Acknowledgments

We wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

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 the Digital Age (BD 2019), CEUR Workshop Proceedings, 2022. Forth-coming.

[5] P. Leskinen, M. Koho, E. Heino, M. Tamper, E. Ikkala, J. Tuominen, E. Mäkelä, E. Hyvönen, Modeling and using an actor ontology of Second World War military units and personnel, in: C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, J. Heflin (Eds.), The Semantic Web – ISWC 2017: 16th International Semantic Web Conference, volume 10588 of *Lecture Notes in Computer Science*, Springer, Cham, 2017, pp. 280–296. doi:10.1007/978-3-319-68204-4_27.

[6] A. Meroño-Peñuela, A. Ashkpour, M. Van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach, F. van Harmelen, Semantic technologies for historical research: A survey, Semantic Web 6 (2015) 539–564. doi:10.3233/SW-140158.

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

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

[9] 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.

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

[11] 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.

[12] M. Koho, P. Leskinen, E. Hyvönen, Integrating historical person registers as Linked Open Data in the WarSampo knowledge graph, in: E. Blomqvist, P. Groth, V. de Boer, T. Pellegrini, M. Alam, T. Käfer, P. Kieseberg, S. Kirrane, A. Meroño-Peñuela, H. J. Pandit (Eds.), Semantic Systems. In the Era of Knowledge Graphs. SEMANTiCS 2020, volume 12378 of *Lecture Notes in Computer Science*, Springer, Cham, 2020, pp. 118–126. URL: https://doi.org/10.1007/978-3-030-59833-4_8.

[13] 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.

[14] M. H. 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.

[15] 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.

[16] 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.

[17] E. Ikkala, M. Koho, E. Heino, P. Leskinen, E. Hyvönen, T. Ahoranta, Prosopographical views to Finnish WW2 casualties through cemeteries and Linked Open Data, in: A. Adamou, E. Daga, L. Isaksen (Eds.), Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II), volume 2014 of *CEUR Workshop Proceedings*, 2017, pp. 45–56.

[18] L. Rietveld, R. Hoekstra, The YASGUI family of SPARQL clients, Semantic Web – Interoperability, Usability, Applicability 8 (2017) 373–383.