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

Mikko Koho\textsuperscript{1,2}, Heikki Rantala\textsuperscript{1,2} and Eero Hyvönen\textsuperscript{1,2}

\textsuperscript{1}Semantic Computing Research Group (SeCo), Aalto University, Espoo, Finland
\textsuperscript{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, which contains rich data about Finland in the Second World War in Linked Open Data format. 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 comparative studies. This paper gives an overview of the possibilities of WarSampo, and presents examples of how WarSampo, and by extension Linked Data more generally, can be used to created various analyses to help with historical research.

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

1. Introduction

WarSampo – Finnish Second World War (WW2) on the Semantic Web \cite{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\textsuperscript{1}. The data is available via an open SPARQL endpoint and a public web portal\textsuperscript{2} for searching, browsing, and analyzing the data through nine different perspectives.

The core dataset of WarSampo \cite{2} is the casualty register \cite{3} of the National Archives of Finland, containing detailed information about all 94,700 people killed in action in Finland during WW2. 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 a person register of 4200 Finnish prisoners of war (POW) \cite{4} and 5600 notable individuals from other data sources (Wikipedia, etc.) \cite{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 for external tools. Just SPARQL 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 950,000 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 enlisted to officer ratio, 2) how the perished soldiers’ places of domicile correlate with their mortality, social class, and place of burial, and 3) how occupation affected the likelihood of surviving from POW camps.

2. WarSampo Knowledge Graph

Linked Data and the event-based CIDOC Conceptual Reference Model (CRM) [11] are used together as a basis for harmonizing various datasets about Finland in the Second World War into a unified KG. Main entity types are persons, military units, death records, prisoner records, events, places, photographs, war diaries, articles, and occupations. The death and prisoner records correspond to structured metadata records of persons from the casualty register and POW register, respectively. These are linked to their corresponding person entities [12].

The casualty register of WarSampo is of great importance in studying Finland in WW2, as it contains detailed information about all of the 95,000 perished soldiers in the Finnish front. In addition to this, WarSampo contains basic information of 5600 notable persons from various sources, and a very detailed person register of all 4200 Finnish POWs.

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, i.e., a large collection of wartime photographs, hand-written digitized war diaries, and war veteran magazine articles. These easily provide further information for people studying for example the war paths of their relatives. The newest 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 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 persons provides a point of access to all of the information about the person contained in WarSampo, making it possible 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 project\(^4\), 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

The faceted search user interfaces of WarSampo Portal provide an easy way for anyone interested in military history to study, explore, and analyze the integrated datasets in the WarSampo KG. An interested citizen can do his/her own data analysis on the data, which would generally often be impossible as detailed historical data tends to otherwise be siloed to Universities and their prominent researchers.

A number of data analytical visualizations can be performed on the WarSampo portal [17, 4]. Figure 1 presents a screenshot of the soldier life paths diagram from the Casualties perspective [17], showing the life paths of 40 soldiers. The diagram shows where the soldiers were born, 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 cemetery of Inari in Ivalo.

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 SPARQL endpoint. 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 certain visualizations can be created for prosopographical analyses using the WarSampo KG.

---

\(^3\)WarSampo Knowledge Graph: https://www.ldf.fi/dataset/warsapo

\(^4\)The WarMemoirSampo project: https://seco.cs.aalto.fi/projects/war-memoirs/en/
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 similar, 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 triples 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 consisting of a subject, a predicate, and an object. 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 order the results based on the dates in ascending order. These results, that consist of dates and numbers, can be visualized,

\(^5\)Yasgui can be used to easily share SPARQL queries. You can see and test the query 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.

e.g., as a line chart as show in figure 2 created with the Yasgui tool [18].

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 .

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 Finish province based on the municipality of residence of each victim. It is easy to see that the Uusimaa Province
3.3. Survival From Prisoner-of-War Camps

The register of the Finnish POWs is much smaller than the casualty register, with 4200 persons, 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. The large group “No HISCLASS code” corresponds mostly to non-specialized workers who worked in various tasks, often seasonally (“työmisc” and “sekatyömies” in Finnish).

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 of the 10 most common occupations in this group is shown in 5. This would suggest that many skilled

---

6You can see and test this query here: https://api.triplydb.com/s/_rLiQGFp
7The 7-class HISCLASS distribution of POWs: https://api.triplydb.com/s/DSwhf3w45
8Occupations of the group "Foremen and medium skilled workers": https://api.triplydb.com/s/arnMtmYEq
workers had better chances of survival than others. As suggested by a collaborating historian, some skilled workers have been valued during their imprisonment as they have been useful in practical tasks, and thus might have received better treatment.

4. Discussion

In this paper we have shown how LOD can facilitate historical research, as exemplified with the WarSampo KG. Metadata that is harmonized and reconciled into ontologies, like the ontology infrastructure of WarSampo, provide enhanced possibilities for information retrieval that,
combined with basic visualizations, 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 effort to master.

To go beyond these 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 to understand what would be interesting to study and how to interpret the produced results and set 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. Using the possibilities fully will require developing a strong ontology infrastructure to which 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


