Reassembling the Lives of Finnish Prisoners of the Second World War on the Semantic Web

Mikko Koho, Esko Ikkala, Eero Hyvönen
Semantic Computing Research Group (SeCo)
Aalto University and University of Helsinki, Finland
HELDIG – Helsinki Centre for Digital Humanities
firstname.lastname@aalto.fi

Abstract
This paper presents first results of a new, ninth application perspective for the semantic portal WarSampo – Finnish WW2 on the Semantic Web, based on a database of ca. 4 460 Finnish prisoners of war in the Soviet Union. Our key idea is to reassemble the life of each prisoner of war by using Linked Data, based on information about the person in different data sources. Using the enriched aggregated data, a biographical global “home page” for each prisoner of war can be created, that is more complete than information in individual data sources. The application perspective is targeted to researchers of military history, to study and analyze the data in order to form new research questions or hypotheses, as well as to public in the large looking for information e.g., about their relatives that were captured as prisoners of war. Employing the faceted search of the application perspective, prosopographical research on subgroups of prisoners is possible.

1 Introduction
Representing biographical texts as Linked Data leads to a paradigm change in publishing biographical collections (Hyvönen et al., 2019): the lives can then not only be read as texts by humans but also be processed and analyzed by computational means (Fokkens et al., 2017; Warren et al., 2016), opening new possibilities in Digital Humanities (Gardiner and Musto, 2015) research for biography and prosopography (Verboven et al., 2007) as well as for data reuse in applications. The same idea of Linked Data can be applied also when biographical data is available in semi-structured or structured form from different data sources: the data about a person can be aggregated, harmonized, and reassembled into a global knowledge graph that gives a more complete picture of the biographee than any individual source alone. Based on the knowledge graph, a biography of the biographee can be generated or alternatively a semi-structured “home page” presenting her/his life. The latter approach was introduced in the semantic portal WarSampo – Finnish WW2 on the Semantic Web1 (Hyvönen et al., 2016), a web service in use in Finland that had 230 000 users in 2018, typically looking for information about their relatives killed in action during the Second World War (WW2).

This paper presents a new, ninth application perspective Prisoners of War to be included in WarSampo. This perspective was created for studying individual persons, documented in a new prisoners of war (POW) database, as well as groups of them for prosopographical analysis. The new data was aligned with and integrated into the WarSampo person data, which is mostly based on the Finnish WW2 casualties of war2 (Koho et al., 2017) database of the National Archives of Finland. The new application perspective enables studying not only individuals but also prosopographical studies of the prisoners using either the whole dataset or subsets of it based on user interest and selections in a faceted search (Tunkelang, 2009) view.

The new prisoners of war data, which extends the person home pages further, has originally been published as a book (Alava et al., 2003). For integrating and publishing the data as a part of WarSampo, it has been further extended, cleaned, and validated by domain experts using, e.g., information from many war-time archives. This paper builds on previous work discussing the Linked Data publication and data model (Koho et al., 2018a), and the data integration challenges (Koho et al., 2018b).

In the following, the underlying data model and data production process is first explained. After this, main functionalities of the application from an end user perspective are explained, as well as the technical implementation. In conclusion, contributions of the work are summarized and contrasted with related work.

2 Data Model and Data
The prisoner data consists mainly of a register of Finnish prisoners of war in WW2, containing a spreadsheet of about 4 460 soldiers, auxiliary forces, and civilians caught by the army of the Soviet Union. Additional spreadsheets contain information about POW camps and hospitals, as well as the primary data sources. The data includes also separate documents about the prisoners of war to provide additional information, such as video interviews, images and archived documents.

The original information sources are mostly various registers in Finnish and Russian archives (Alava et al., 2003). Information in different sources can be contradictory, hence it is important to preserve the data source for each individual pieces of information. A formatting was agreed upon to allow multiple values with source information already

---

1This semantic portal was released in 2015 and is in use at https://sotasampo.fi/en/. More information about the project is available at home page https://seco.cs.aalto.fi/projects/sotasampo/en/.
2http://kronos.narc.fi/menehtyneet/
in the original spreadsheet that the domain experts worked on. The data formatting evolved as a collaboration between the domain experts maintaining the original dataset, and the WarSampo team of Linked Data experts. Also other agreements on the spreadsheet structure were needed: 1) separation and cleaning of values that will be linked to the WarSampo domain ontologies, 2) local identifiers for entities that are used in multiple spreadsheets, and 3) how to express partially or completely missing information.

The WarSampo data service and semantic portal was chosen as the primary data publication platform by the stakeholders, which include the National Archives of Finland, and the Association for Cherishing the Memory of the Dead of the War.

21 Prisoners of War as Linked Data

The WarSampo Linked Open Data infrastructure is built to support integrating new datasets into WarSampo, by extending both the data model and the data content. The data is published openly online for everyone to use. The WarSampo web portal then provides different perspectives to the interlinked datasets, as customized web applications. New perspectives can be added to provide views to new datasets, or to show new features of the existing data.

In Linked Data (Heath and Bizer, 2011), data is presented as RDF graphs, and all resources in the data have unique identifiers. This enables identifying and sharing common resources, e.g. people, places, and military ranks between the datasets, thus creating an interlinked knowledge graph. A simple primary data model is used for the prisoner records, in which one prisoner record corresponds to one row in the source spreadsheet, with each column mapped to a distinct property. So all of the personal information about each captured individual is contained in the prisoner record, resembling the data model of the WarSampo death records (Koho et al., 2017). The properties and classes of prisoner records and death records have been harmonized using the dumb-down principle of Dublin Core, i.e., by using shared super-properties and super-classes where applicable. By mapping columns directly to properties, the data can be shown to the end user in an intuitive way, resembling the original spreadsheet.

WarSampo uses the CIDOC Conceptual Reference Model (CRM) as the harmonizing data model. Prisoner records are modeled as instances of the CRM document class E31_Document.

In addition, this data is then used to create CIDOC CRM descriptions of the actual persons and events, when appropriate. WarSampo person instances (Leskinen et al., 2017) in the actor ontology are enriched using the prisoner records. New person instances are created for persons that do not already exist in the ontology, which is the case for most of the war prisoners. The prisoner records then document the person instance through the CRM property P70_documents.

22 Data Conversion

A data processing pipeline was developed to integrate the POW data into WarSampo linked data infrastructure. The pipeline handles data transformation, validation, linking, and harmonization.

The pipeline transforms the spreadsheets into RDF, mapping the spreadsheet columns to RDF properties, with possibly multiple values per property, and containing annotations for primary information sources. Automatic probabilistic entity linking processes then link the records to WarSampo domain ontologies of military ranks, units, occupations, people, and places. Original literal values are also retained as separate properties.

The original POW register is maintained in spreadsheet format, which can be easily integrated into WarSampo with our automated transformation process when the spreadsheet is updated, provided that the structure stays the same. Also if the linked domain ontologies are updated, the whole integration process can be redone to account for the changes in the probabilistic entity linking.

The cell formatting is validated during the data transformation process. Also other simple data validation rules are applied to find anomalies during data conversions. Some parts of the data had to be left out of the online data publication due to privacy legislation. This is done automatically based on the date when a person has died. In some cases, the individuals need to be anonymized to keep their identity hidden.

23 Interlinking within WarSampo

Matching the people in the prisoner records to those already existing in the WarSampo actor ontology is one of the most challenging aspects of the data transformation pipeline. The data model and contents are different, and many pieces of personal information can be missing on both sides. In the first results of the person linking, we were able to link 1417 prisoner records to existing WarSampo person instances, corresponding to 32% of all prisoner records (Koho et al., 2018a). The person linking uses probabilistic record linkage (Gu et al., 2003; Gregg and Eder, 2019) with a machine learning approach, in which each POW’s information is compared with the information in the WarSampo person instances to find matches that have high enough similarity. Initially the comparisons were weighted based on domain knowledge, which was then iterated to work better, and manually curated to serve as training data for the machine learning approach. This approach can adapt to data changes on both sides in the record linkage, without having to manually inspect the linking results and adjust the weights.

New person instances are then created for the unlinked POW records and added into the actor ontology. With the probabilistic record linkage, it is possible that a record is not mapped simply because there is not enough information about either the POW record, or the person instance, to create a mapping between them. Modifying the informa-

---

3http://dublincore.org/usage/documents/principles/
4http://cidoc-crm.org

5Source codes for data conversion and linking are available online: https://github.com/SemanticComputing/WarPrisoners.
tion in either the POW data or in the actor ontology means that the whole record linkage process should be redone. Other information is also linked to WarSampo domain ontologies. Of military ranks, 99% were linked to the WarSampo military ranks domain ontology. Of military units, 90% were linked to military units in the actor ontology. The missing cases should be manually inspected, and if it is a lacking in the domain ontology, the domain ontology should be updated to handle the unlinked values, else the POW data should be updated to use the terms in the domain ontology.

Domain ontologies differ from each other by nature. For example, covering and disambiguating all military ranks is clearly a simpler task than performing the same task with all wartime places. In general, it is not realistic to assume that the domain ontologies completely cover their domain. Other information still to be linked to WarSampo domain ontologies are war-time municipalities. More accurate place information could also be linked, but due to the ambiguous nature of the names, this would lead to a high level of error, based on initial experiments.

The created Linked Data stores source information when present in the original data. There are many ways of presenting this kind of provenance information in RDF (Hartig, 2009; Zhao et al., 2010). The approach used with the prisoners of war dataset is storing source information using RDF reification with the DCMI Metadata Terms\(^6\) property source.

### 24 Biographical Data

Basic personal information of each person in the data contains columns like first and last names, dates of birth, return from captivity, and death, municipality of birth, domicile and death, and occupation, marital status, and number of children. These enable building some understanding of the life of the person before the war, and in case of survivors, also after the war.

Structured information is also gathered of the events of going missing and being captured, like place and time. Biographically interesting information is also given as prose about being captured, cause of death and burial place, and other information. These all are structured to contain the information source, and can often contain different pieces of information from different sources. Information of confiscated possessions and their estimated value sheds light to what kind of valuable personal possessions a person had. Information is given also of the occurrence of a person in Soviet war propaganda magazines or fliers, either in pictures or text.

### 3 Prisoners of War in the WarSampo Portal

A new application perspective was created into the WarSampo portal for studying, exploring and analyzing the prisoners of war dataset as a whole. Also the existing WarSampo Persons perspective, which generates a “home page” for each person in the WarSampo knowledge graph, was extended to show possibly contradictory data originating from multiple sources (e.g. death records, prisoner records, Wikipedia). The Prisoner perspective application is open-source, and available online.\(^7\)

### 31 Biographical View in the Persons perspective

The WarSampo Persons perspective offers a general search of people in the WarSampo knowledge graph. Each person is provided with a biographical view, a “home page”, that reassembles the biographical knowledge of the person from the WarSampo datasets, into a structured format.

Figure 1 shows an example of a soldier’s “home page”, in which the information is combined from a prisoner record and a death record. The left side of the page contains a person selector and a text box for filtering the people by name. The details of a selected person is displayed on the right. Information usually exists from birth to death, with a clear and understandable focus on the war-time events. One property (e.g. occupation) may contain multiple values. In order to make the biographical view as transparent as possible, all values have been supplemented with a link to the information source.

The values that have been linked to WarSampo domain ontologies are shown as links to corresponding “home pages”. The idea here is that the WarSampo semantic portal acts as a customized graphical RDF browser, which makes it possible for the user to find surprising connections between the individual resources of the WarSampo knowledge graph.

### 32 Prosopographical Prisoners Perspective

The Prisoners perspective is based on the previously released Casualties perspective (Koho et al., 2017). The main design principle of these perspectives is to target one core class of WarSampo knowledge graph (e.g., prisoner record) and provide the user with a faceted search (Tunkelang, 2009; Oren et al., 2006) interface, which initially renders a result set that contains all instances of the target class as a paginated table. This way we ease off the “blank search field problem”, where a new user does not know what kind of query terms should be used for meaningful results. The initial result set can be narrowed down by using various facets (e.g., military unit or prison camp).

Figure 2 shows a part of the Prisoners perspective user-interface. Facets are presented on the left of the user-interface. The number of hits (instances of the target class) produced by each facet value is calculated dynamically and is shown in parenthesis. Facet values leading to an empty result set are hidden. To reduce unnecessary data fetching, most of the facets are disabled by default. They can be activated by clicking the plus sign on the facet header. The facets are name, date of being captured as a POW, date of death, military unit, military rank, POW camps where the person has been, occupation, marital status, number of children, birth municipality, place of being captured, and place of death.

The results are displayed on the right side of the user-interface. The result set, based on the facet selections, can be shown as a table, or shown with 3 different visualizations:

\(^6\)http://dublincore.org/documents/dcmi-terms/

\(^7\)https://github.com/SemanticComputing/prisoners-demo
Figure 1: The Persons perspective showing part of a person’s home page.
1. a distribution chart over a selected property, with property choices: military rank, military unit, occupation, number of children, birth municipality, municipality of residence, place of being captured, and place of death,

2. an age distribution chart at the time of capturing,

3. a sankey diagram of soldier life paths based on known geographical locations at different times, starting from the municipality of birth, and ending to the municipality of death.

The results display mode can selected using the button on the top bar. In Figure 2, the results are displayed as a table, with each row corresponding to a single prisoner record, with several key properties mapped to separate columns. Figure 3 shows the age distribution of all soldiers whose rank is private at the time when they have been captured as a prisoner of war. Figure 4 shows the military rank distribution of the soldiers that were born in Helsinki.

The common usage scenario of the average user is to search for information about their relatives who have participated in the war. This can be achieved most easily with the table view of results and using the different facets, and mostly the name facet, where a person can search with just a part of the name to get all the results containing that. Another way to find relatives, who historically are often situated in the same region, is to filter the results with the birth municipality facet.

Another usage scenario is studying and analyzing the data by a historian or an interested citizen. The facets already provide distributions of the facet values, with the number of hits after each value. When a selection is made in one of the facets, all of the facets are updated to show the distribution of values with that selection. Further analysis can be done with the various visualizations of the facet results. New visualizations, e.g. locations of prisoners camps on a map, can be added rather easily to the application, and the existing ones extended as needed.

4 Implementation of the Prisoners Perspective

The Prisoners perspective is an AngularJS8 web application, which consists of several modules. The facet func-

8https://angularjs.org/
Facet functionality is implemented using SPARQL Faceter⁹, a module that provides:
- a set of directives that work as configurable facets
- a service that synchronizes the facets
- a service for updating the URL parameters based on facet selections, and retrieving the facet values from URL parameters
- a service for retrieving SPARQL results based on facet selections

For querying the SPARQL endpoint, mapping the SPARQL results into JavaScript objects and paging the results, we...
have developed another general module\(^\text{10}\) that is being used across the WarSampo semantic portal. The SPARQL Faceter module offers a configurable query template for retrieving the results based on the facet selections. In addition to the paginated table, powered by the ngTable\(^\text{11}\) directive, we have implemented several reusable visualization directives for displaying the results. For the Prisoners perspective, a new sankey visualization directive was built using Google Charts.\(^\text{12}\)

5 Discussion

This paper presented first results of publishing the prisoners of war dataset as part of WarSampo. The POW data contains sensitive information about the individual citizens, some of whom are still alive. The publication of the data has been delayed due to the evaluation as to what information can be legally published about the individuals, and what needs to be hidden. The dataset and new portal is expected to be finally published in 2019.

Maintaining interlinked datasets and domain ontologies present new challenges (Auer et al., 2012; Maedche et al., 2003), as changes is one part need to be accounted for in other interlinked parts. The Linked Data Environment is not yet mature enough to have easy-to-use tools for non-technical people to use for editing and maintaining interlinked data. Hence, the POW data is still maintained using the spreadsheet with agreed upon formatting and structuring, which can then be re-integrated easily into WarSampo. The Linked Data approach requires tighter co-operation with the domain experts and data publishers, especially in the creation phase of historical information (Boonstra et al., 2004), than more traditional data publishing ways. As Linked data, it is however possible to create an understanding of the whole of the war, by combining information from several datasets together, which would not be easy by studying the individual datasets directly.

The historical occupations in the WarSampo datasets have recently been harmonized into a manually curated ontology, AMMO (Koho et al., 2019), to which the prisoner records will be linked, in addition to death records and person instances. Using the AMMO ontology enables studying the prisoner records using new facets, such as social class, and field of work. Integrating data into a Linked Data infrastructure is more laborious than simpler ways of publishing the data as an independent data object, which does not communicate with other datasets. However, the result of the integration is an interlinked knowledge base, where the interlinked graphs enrich each other, creating a whole that is greater than the sum of its parts.

Acknowledgements

Reijo Nikkilä, Tiia Moilanen, and Pertti Suominen of The National Prisoners of War Project worked on the data as domain experts.

Our work was funded by the Association for Cherishing the Memory of the Dead of the War\(^\text{13}\), Open Science and Research Initiative\(^\text{14}\) of the Finnish Ministry of Education and Culture, the Finnish Cultural Foundation, and the Academy of Finland.

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

6 References


Eileen Gardiner and Ronald G. Musto. 2015. The Digital Humanities: A Primer for Students and Scholars. Cambridge University Press, New York, NY, USA.


\(^{10}\)https://github.com/SemanticComputing/angular-paging-sparql-service

\(^{11}\)https://github.com/esvit/ng-table

\(^{12}\)https://github.com/ angular-google-chart/angular-google-chart

\(^{13}\)http://www.sotavalnajat.net/in_english

\(^{14}\)http://openscience.fi/


