Building a Linked Open Data Portal of War Victims in Finland 1914–1922

Heikki Rantala[1], Ilkka Jokipii[2,3], Mikko Koho[1], Esko Ikkala[1], Jouni Tuominen[1,2], and Eero Hyvönen[1,2]

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

Abstract. This paper presents results from a project that publishes data about the war victims in Finland in 1914–22 as a Linked Open Data service and a portal of tools called WarVictimSampo 1914–22 to explore and analyze the data. At the same time, the data is extended with new information and cleaned from mistakes when found. The project is based on the database War Victims of Finland 1914–22 ("Sotasurmat 1914–22") of the National Archives of Finland and related data compiled during the project. This database contains some 40000 death records, most of which are due to the Finnish Civil War but some are related to the First World War and the Kindred Nations Wars. In addition to the war victims, our data includes information about 1200 battles of the Civil War and various prisoner camps. A key novelty of WarVictimSampo 1914–22 is the integration of ready-to-use Digital Humanities tooling with the data service, which allows, e.g., studying information about wider prosopographical groups in addition to individual victims. We show by examples how using the tools of the portal the data can be better explored and analyzed for research and education purposes.

Keywords: Linked Data, Semantic Web, War History.

1 Introduction

This paper presents first results of a project[4] where the database of Finnish War Victims 1914–22 is updated and converted to Linked Data. In addition new and better tools for accessing the data are created. The aim of the project is to help in studying the war related deaths between the years 1914 and 1922. The Finnish Civil War is the main focus point for the project: 93 percent of the deaths recorded in the database are related to the Finnish Civil War and its aftermath in 1918, and the rest to the First World War and the Kindred Nations Wars (Heimosodat). The original data was recorded in 1999–2003 using a custom-made tool and it includes 39931 records. Fig. 1 shows the daily distribution of death dates in the data. [15]

There is an old web application[5] in use for exploring the data with simple search functionality and a homepage for each person. The person's homepage shows basic

[4] https://seco.cs.aalto.fi/projects/sotasurmat-1914-1922/
[5] http://vesta.narc.fi/cgi-bin/db2www/sotasurmaetusivu/main

information about the victim, but a lot of information is not shown although it would be available in the underlying database. The end users of the system have deemed the search interface fairly inflexible with too few options to choose from. Also some means of exporting the data from the database has been asked for.

In order to provide the war historians and the public in the large a more versatile and complete open access to the data we created a new data service, based on the FAIR principles[6], and a set of web applications on top of it. The new applications[7] include not only tools for searching and browsing the data, but also tooling for Digital Humanities (DH) [11, 3] research to analyze and study the data. Researchers will also be able to download the data filtered by faceted search if needed in CSV form for re-using it in their own tools and applications.

As for the methodological approach, Semantic Web (SW)[8] and Linked Data (LD) technologies [4] were chosen as well as the "Sampo" publishing model[9] [6]. In this way cross-disciplinary and cross-organizational data integration is made easier by representing and processing data using shared semantics and ontologies. This is useful for semantic data interoperability, enrichment, validation, exploration, visualization, and knowledge discovery.

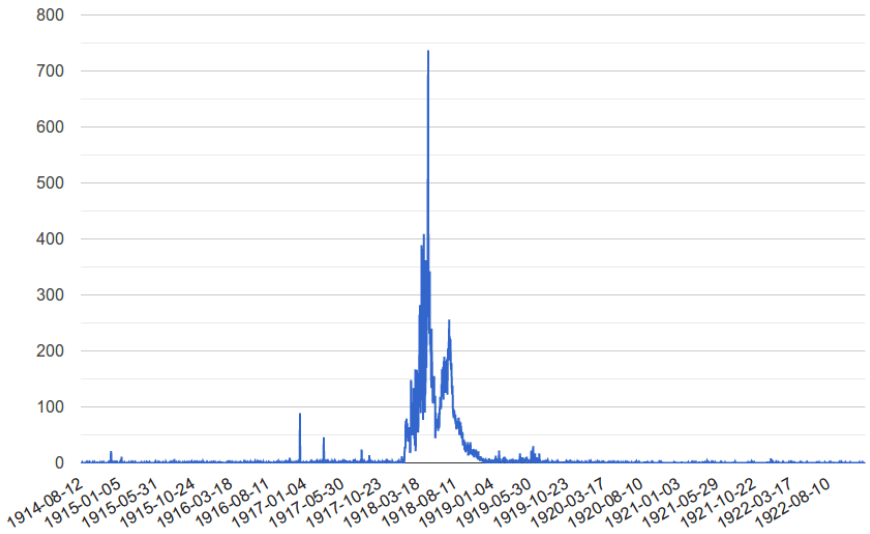

Fig. 1. Distribution of death dates as shown in WARVICTIMSAMPO 1914–22.

[6] Findable, Accessible, Interoperable, and Re-usable, cf., https://www.go-fair.org/fair-principles/

[7] https://sotasurmat.narc.fi/

[8] http://www.w3.org/standards/semanticweb/

[9] https://www.europenowjournal.org/2019/09/09/linked-data-in-use-sampo-portals-on-the-semantic-web/

The paper is structured as follows. First, the data underlying our work is presented. After this, the data conversion into Linked Data is in focus. To illustrate the possibilities of the new system for Digital Humanities research, tools for data exploration, analysis, and visualization are then presented. In conclusions, related works are discussed and a plan for the future is outlined.

2 The Data

The death records contain basic information of the people (e.g., name, place of birth, date of birth, date of death), socioeconomic information (e.g., occupation, marital status), and war related information (e.g., military rank, military organization, time of imprisonment). On average each death record is based on more than three different data sources. The most important sources are Clerical statistics, Parish Records, Social Democratic Party Records, and the Prisoners of War Archive. [15]

After publishing the original War Victims Database, the National Archives of Finland (NAF) has received lots of corrections and new data from active citizens and from the research community. NAF therefore started a project in late 2018 which aimed to incorporate this information into the old data. Despite the large average amount of data sources, the old data has many records that barely give basic information of the victims, and the new project is also aiming to enrich the existing records. Our work in the follow-up project also included incorporating new sources into the database, such as the Archive of Committee to Support the Widows of 1918, which sheds more light on the existing records and also brought up 1590 previously unknown war-related deaths.

The conversion into Linked Data (LD) was made from CSV files provided by NAF. The data was provided in three files: one core table including the main information about each person, one for additional information, and one for the sources. The new LD data model[10] was made generally similar to the original model, and uses the same original codes for the people. A main difference is that in the new version all entities have a unique URI that makes it easy to reference them internally and in external datasets and applications.

A major part of the conversion was made using the RML mapping language[11]. Parts of the conversion were also made using the Python RDFLib library and SPARQL CONSTRUCT queries. The original data is rich and has a lot of different ways in which various pieces of information can be related to a person. This means that the data model needs to be relatively complex to express everything. A major goal in our work has been to create a conversion into LD using an easily understandable model. Standard vocabularies, such as SKOS and CIDOC CRM, have been used where applicable.

One interesting and non-trivial question encountered in the work was how to express the dates in a machine understandable form when the exact date is unknown. For example, only the month of the birth date might be known. This might be expressed in the data as a literal value such as "?.3.1899". In our case, a solution based on CIDOC CRM is used. Dates are expressed as time span resources that have properties

[10] See http://www.ldf.fi/dataset/siso for the documentation of the data model and access to the data.

[11] http://rml.io

"begin_of_the_begin" and "end_of_the_end" modeling the earliest and latest possible birth date, respectively. Therefore the birth date would be a time span of "1.3.1899-31.3.1899". When using this kind of time spans, it is important to note the difference between a date that may be within a time span and a date certainly is within a time span. For example, a researcher might be interested in people born between 7.3.1899 and 12.3.1899. The birth date of the above example would possibly be within this time span, but not certainly. Allowing a wide interpretation in search that gives the user all the time spans that possibly fit the desired time span is usually the best choice, but it maybe be confusing, too. If we only know that a person was born in the 19th century, then he may have been born between 7.3.1899 and 12.3.1899. However, if this person is included in a search result of people born between 7.3.1899 and 12.3.1899, the user might consider such a search result weird, because the probability in this case of being born on these few days is very low. This means that extra attention must be given to explain to the user the limitations of the data and related probabilities.

The linked data approach makes it easy to link the data to and from outside sources. In our case, for example, a mapping has been made to the AMMO [9] ontology that defines historical occupations in Finland, interlinked with the international HISCO classification [14]. AMMO ontology was partially made using the War Victims data. This ontology includes information about the social status of occupations and may be useful in researching the Civil War victims.

The data in LD form is loaded into a SPARQL endpoint hosted at the Linked Data Finland platform[12] [5], where it can be queried using the SPARQL query language. Our web application makes queries to this publicly open endpoint, and also a researcher can query the database for her own purpose.

What is mainly modeled in our data are not people, but instead "death records". A person instance could later be created based on a death record, as in WarSampo [8]. In the LD model, every fact in a record is a separate resource with a URI. This makes it easy, among other things, to attach a source for each fact. This is conceptually close to the idea of RDF reification. The information resources have properties for the type of the information, value, reliability of the information, sources, and other metadata. Death records with simple table-based datamodel are generated from the information resources. The death records include only one, the most reliable, value for only the most important types of information. These death records make the faceted search more efficient and allow easy generation of the most important data as a simple table.

3 Web Application

A user interface was developed to allow different user groups to access the data easily. Potential user groups include researchers, students, and the wider public interested in either the Finnish Civil War in general or the fates of their relatives. It is obvious that school children can't be expected to create their own SPARQL queries to query the data. That can be too much to ask from even a researcher of history. On the other hand, even a researcher who is able to create her own queries may find it useful to have an

[12] http://www.ldf.fi/dataset/siso

easy way to explore the data and to create simple visualizations quickly. Visualization are hopefully useful for both finding new data and educating the public about history. These tools should not be expected to fully replace other research. They should only be used to spot interesting things that require more careful research.

The user interface has been implemented as a full stack JavaScript web application. It consists of a client based on React[13], and a NodeJS[14] backend build with Express framework[15]. The state of the application (e.g., the user's facet selections) is maintained using Redux[16]. Asynchronous data fetching is carried out by Redux Observable[17] functions, which make API calls to the NodeJS backend. The data flow in the backend consists of three steps: 1) the relevant part of the state of the application is converted into a corresponding SPARQL query, 2) the SPARQL query is run against the configured endpoint, and 3) the results of the SPARQL query are mapped into JavaScript objects and sent to the client. The backend also supports CSV result format, which skips the result mapping step.

The search results can be rendered using custom-built React components including a table based on Material-UI[18], a map based on Leaflet[19], and chart visualizations based on Google Charts[20]. New components for displaying the results can be easily added using the current modular application architecture.

4 Exploring the Data

The web application is built around the concept of faceted search [13]. With faceted search, the user can easily narrow the search step by step by making selections based on predetermined orthogonal hierarchies of property values called facets. Facets also show the number of available items with each possible selection. This allows the user to immediately see the number of solutions of each possible selection. For example, the "Death Province" facet might show the number 8898 for the selection "Häme Province" in the hierarchy and the number 918 for the selection "Mikkeli Province". Hence the user can observe that there are far more people in the data who died in the Häme Province than in the Mikkeli Province. In this way, the user is guided for making informed filtering selections during search and never ends up in a "no hits" dead end.

Combined with selections on other facets like occupation, party, and age, the user may also draw interesting conclusions by observing the hit distributions on the facets. Thus faceted search can be used to find individuals, such as relatives, that fit certain criteria, but it can also be used to find information about the distributions of different kind of the casualties. This kind of search paradigm is an example of exploratory search [10].

[13] https://reactjs.org
[14] https://nodejs.org/en
[15] https://expressjs.com
[16] https://redux.js.org
[17] https://redux-observable.js.org
[18] https://material-ui.com/components/tables
[19] https://leafletjs.com
[20] https://react-google-charts.com

The application currently includes two main application perspectives for exploring the War Victims data: 1) The main perspective is based on searching and exploring the casualties, currently 39 931 death records. 2) There is also a perspective based on the battles of the Finnish Civil War, currently 1182 geo-coded battles. Other views may be added later in the same way as in other "Sampo" series semantic portals[21].

For the both application perspectives there are multiple tabs to view the data in different ways. For example, in addition to the list view (table) of casualties that includes the names and basic information about him/her, there are different visualization options available for inspecting the data filtered using faceted search, such as pie and line charts and map views. There is also a tab to download the current selection of entities in CSV form. The search functions and the visualizations can be useful in educating people about the Civil War, but can also be useful for researchers. The aim is to allow the user to explore the data and find useful or surprising information and patterns in the data.

For example, Fig. 2 illustrates how the pie chart tab is used to visualize different properties of the casualty data. The user has made two facet selections for filtering the data, Party=Reds and Gender=Woman, using the facets on left. The pie chart visualization based on the facet "Registered Municipality" allows one to see easily that most of the Red women killed were registered to a relatively small town of Sääksmäki. This may not bring new knowledge as such, but just seeing the visualization may help to understand the Civil War in a new way and ability to easily generate such visualizations may help in finding interesting phenomena for further study. In the pie chart visualization the municipalities with small amount of victims are grouped automatically together for clarity.

There is also a line chart visualization with options for age, birth year, and the death date. The age visualization also calculates the average and median age of the selected people. This allows, for example, to easily see that people on the Red side who were registered in the Viipuri Province were, for some reason, clearly older than the Red people in the data in general, and those registered in the Turku Province clearly younger. The Red casulties have a median age of 28, while those registered in the Viipuri Province have a median age of 31, and those registered in the Turku Province have a median age of 26. Therefore the Red people in the data registered in the Viipuri and Turku Provinces have roughly a difference of five years in their median age. The application can't directly answer why this is the case, but it can hopefully raise issues like this for which researchers will need to find answers.

5 Related and Future Work

WarVictimSampo is a follow up project of WarSampo [8] that uses LD to present and publish information related to the Second World War in Finland, including death records. The model of WarSampo was used as a reference for this project. However, the research focus in WarSampo has been on issues related to heterogeneous, distributed data harmonization, linking, integration, and LD cloud maintenance[22]. In contrast, the

[21] https://www.europenowjournal.org/2019/09/09/linked-data-in-use-sampo-portals-on-the-semantic-web/

[22] https://seco.cs.aalto.fi/projects/sotasampo/en/

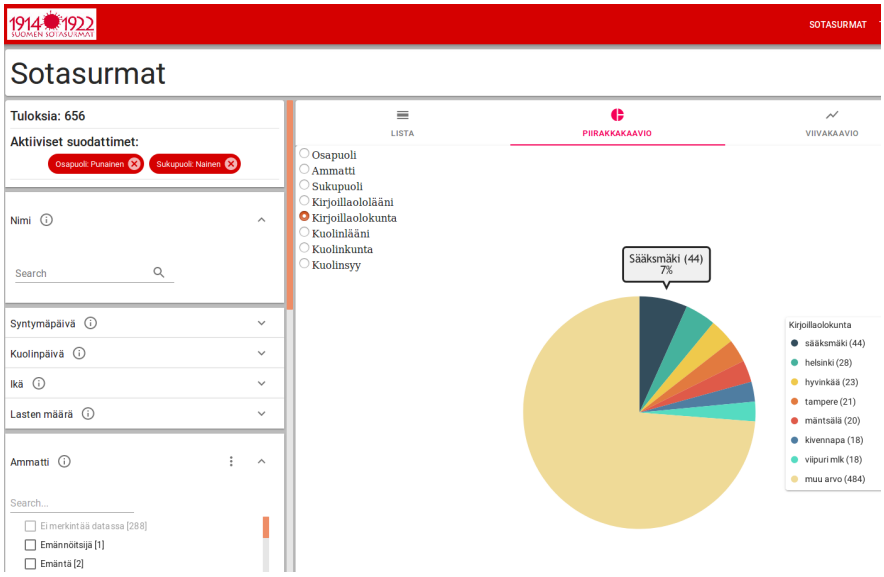

Fig. 2. Pie chart view of the user interface.

data model in WarVictimSampo is much simpler (table-based) and the data are more detailed and systematic, allowing more easily Digital Humanities research based on data analysis and knowledge discovery, as exemplified in the paper. The novelty of WarVictimSampo lays in the idea of developing new data-analytic tooling for research in war history, as well as in creating, cleaning, extending, and publishing the former War Victims 1914–22 database for open use on the the Semantic Web.

There have been several projects publishing linked data about the World War I on the web, such as Europeana Collections 1914–1918[23], 1914–1918 Online[24], WW1 Discovery[25], Out of the Trenches[26], CENDARI[27], Muninn[28], and WW1LOD [12]. In addition to WarSampo, there are a few works that use the Linked Data approach to WW2, such as [2, 1], Open Memory Project[29] on holocaust victims, and the Dutch project Netwerk Orloogsbronnen[30].

In the future, we plan to develop WarVictimSampo into a "third generation" semantic portal, based on Artifical Intelligence, where the machine itself could automatically search for interesting "serendipitous" phenomena like those in the examples of this paper, and perhaps even find explanations for them [7].

[23] http://www.europeana-collections-1914-1918.eu
[24] http://www.1914-1918-online.net
[25] http://ww1.discovery.ac.uk
[26] http://www.canadiana.ca/en/pcdhn-lod/
[27] http://www.cendari.eu/research/first-world-war-studies/
[28] http://blog.muninn-project.org
[29] http://www.bygle.net/wp-content/uploads/2015/04/Open-Memory-Project_3-1.pdf
[30] https://www.oorlogsbronnen.nl

Acknowledgements

Thanks to Päivi Happonen, Vili Haukkovaara, Markku Mäenpää, and Jarmo Nieminen for fruitful discussions and collaboration in the project. Our research was funded by the Prime Minister's Office. Thanks to CSC – IT Center for Science, Finland, for computational resources.

References

1. de Boer, V., van Doornik, J., Buitinck, L., Marx, M., Veken, T.: Linking the kingdom: enriched access to a historiographical text. In: Proc. of the 7th International Conference on Knowledge Capture (KCAP 2013). pp. 17–24. ACM (2013).
2. Collins, T., Mulholland, P., Zdrahal, Z.: Semantic browsing of digital collections. In: Proceedings of the 4th International Semantic Web Conference (ISWC 2005). pp. 127–141. Springer–Verlag (November 2005).
3. Gardiner, E., Musto, R.G.: The Digital Humanities: A Primer for Students and Scholars. Cambridge University Press, New York, NY, USA (2015).
4. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Morgan & Claypool, Palo Alto, California (2011), http://linkeddatabook.com/editions/1.0/
5. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: The Semantic Web: ESWC 2014 Satellite Events. pp. 226–230. Springer–Verlag (May 2014).
6. Hyvönen, E.: "Sampo" model and semantic portals for Digital Humanities on the Semantic Web. In: Proc. of the Digital Humanities in the Nordic Countries (DHN 2020). CEUR WS Proceedings (2020), forth-coming.
7. Hyvönen, E.: Using the Semantic Web in Digital Humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. Semantic Web – Interoperability, Usability, Applicability (2020), in press.
8. 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).
9. Koho, M., Gasbarra, L., Tuominen, J., Rantala, H., Jokipii, I., Hyvönen, E.: AMMO Ontology of Finnish Historical Occupations. In: Proceedings of the The First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH'19). vol. 2375, pp. 91–96. CEUR Workshop Proceedings (June 2019), vol 2375.
10. Marchionini, G.: Exploratory search: from finding to understanding. Communications of the ACM **49**(4), pp. 41–46 (2006).
11. McCarty, W.: Humanities Computing. Palgrave, London (2005).
12. Mäkelä, E., Törnroos, J., Lindquist, T., Hyvönen, E.: WW1LOD: An application of CIDOC-CRM to World War 1 linked data. International Journal on Digital Libraries **18**(4), 333–343 (nov 2017). https://doi.org/10.1007/s00799-016-0186-2
13. Tunkelang, D.: Faceted search. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool, Palo Alto, California (2009).
14. Van Leeuwen, M.H.D., Maas, I.: HISCLASS: A historical international social class scheme. Leuven University Press (2011).
15. Westerlund, L. (ed.): Sotaoloissa vuosina 1914–22 surmansa saaneet. Tilastoraportti. Valtioneuvoston kanslia (2004).