Second World War on the Semantic Web: The WarSampo Project and Semantic Portal

Eero Hyvönen, Jouni Tuominen, Eetu Mäkelä, Jérémie Dutruit, Kasper Apajalahti, Erkki Heino, Petri Leskinen, Esko Ikkala

Semantic Computing Research Group (SeCo), Aalto University, Finland
http://www.seco.tkk.fi/, firstname.lastname@aalto.fi

Abstract. This paper initiates and fosters work on publishing Linked Open Data about the Second World War. It is argued that the heterogeneous, distributed data about the international world war history makes a promising use case for semantic technologies. We hope that by making war data openly available we can learn from the past and promote peace.

1 Publishing Linked Open Data about War History

According to Georg Wilhelm Friedrich Hegel "we learn from history that we learn nothing from history". Hopefully this is not the case for the Second World War (WW2), now that fighting has started again even within the borders of Europe in Ukraine. One way to promote peace is to make reliable data about the war openly available for everybody to learn. WarSampo is a project and semantic portal that aims at this goal by publishing large heterogeneous sets of data about the WW2 in Finland as Linked Open Data (LOD). Application demonstrators are built that provide different perspectives in war history, for both historians and the public. The data covers the Winter War 1939–1940 against the Soviet attack, the Continuation War 1941–1944 where the occupied areas of the Winter War were temporarily regained, and the Lapland War 1944–1945, where the Finns pushed the German troops away from Lapland.

WarSampo[1] is the next step in our series of "Sampo" portals based on Linked Data, including CultureSampo[2] [9], BookSampo[3], and TravelSampo[4] and continues our earlier works on modeling the First World War [6,8]. The project started in autumn 2014 and is finished in 2017, by the centennial of Finland's independence.

2 Data, Metadata Models, and Ontologies

Data The project deals initially with the datasets presented in Table 1. The casualties data (1) includes data about the deaths in action during the wars. War diaries (2) are digitized authentic documentations of the troop actions in the frontiers. Photos and films

[1] http://www.sotasampo.fi
[2] http://www.kulttuurisampo.fi
[3] http://www.kirjasampo.fi
[4] http://www.seco.tkk.fi/projects/subi

Dataset	Name	Providing organization	Size
1	Casualties of WW2	National Archives	93,000 death records
2	War diaries	National Archives	23,000 war diaries of troops
3	Photos & films	Defence Forces & Military Museum	160,000 photos & films
4	Kansa taisteli magazine articles	Bonnier & The Assoc. for Military History in Finland	3,360 articles of veteran soldiers
5	Karelian places	National Land Survey	30,000 places of the annexed Karelia
6	Karelian maps	National Land Survey	War time maps of Karelia
7	Audio & films	National Broadcasting Company YLE	250 recordings and films

Table 1. Central datasets to be linked in WarSampo.

(3) were taken during the war by the troops of the Defense Forces. The Kansa taisteli magazine (4) was published in 1957–1986; its articles contain mostly memories of the men that fought on the fronts. Karelian places (5) and maps (6) cover the war zone area in pre-war Finland that was finally annexed by the Soviet Union. YLE's audio and film material (7) ("Living Archive") was recorded during the war, or is related to it.

Metadata Models CIDOC CRM[5] is used as the harmonizing basis for modeling data, with events providing the semantic glue for data linking [3]. Our data model for WW1, presented in [8], is used as the metadata model to start with.

Domain Ontologies The data is annotated using a set of domain ontologies, including: 1) an ontology of the troops and their hierarchies, 2) persons with their ranks and roles, 3) place ontology of historical places, 4) event ontology of battles, politics, and other war time incidents, 5) an ontology of time periods, 6) ontology of weapons, 7) ontology of vessels, and 8) a subject matter ontology. For 1–7 we have harvested named entities from the datasets, given them URIs and labels and some initial structure, as needed in our initial demos (discussed below). However, ontology modeling and development is still underway. A challenge of the actor ontologies, for example, is modeling the changes: names and positions of the troops as well as the roles of the personnel in the army change frequently (e.g., promotions of persons and changes in troop leadership) and have to be conditioned on time. For 8, the KOKO ontology, a center piece of the Finnish ontology infrastructure [4], is used.

3 Applications: Perspectives to War History

The data and ontologies are published using SPARQL endpoints that form the basis of the WarSampo semantic portal and its applications. The idea of the portal is to provide a variety of different kind of perspectives to war data, represented on different tabs. Most

[5] http://cidoc-crm.org

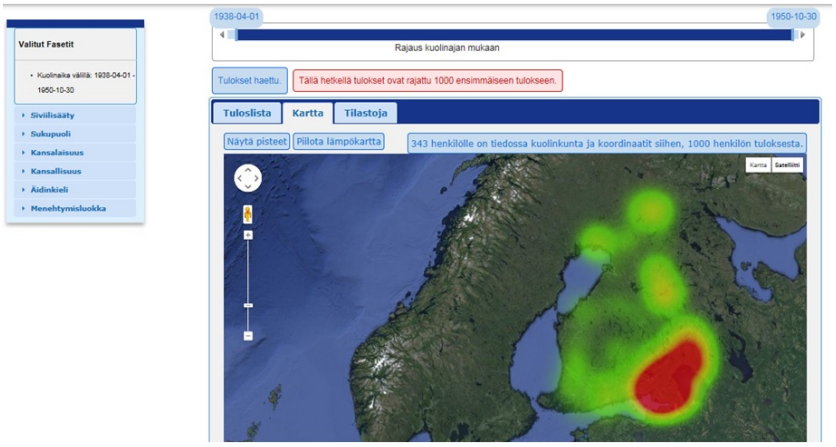

Fig. 1. A heat map illustrating death counts on the map in WarSampo.

datasets will have their own perspective, where the user can first search data of interest and then get linked data related to the resources found. The perspectives enrich each other via Linked Data.

Initial prototypes for two perspectives have already been implemented: one for the war casualty data and one for the Kansa taisteli magazine articles. Fig. 1 depicts the user interface for the casualty data of 93,000 death incidents, with 6 facets on the left (marital status, gender, citizenship, nationality, mother tongue, and death category). On the top, an interactive timeline for the time facet is shown and below it there is a heat map illustrating the death counts on the maps during the selected time interval. Later on, death records will be enriched with links to, e.g., war diaries related to the dead person's troop, related photos, and articles. The second demonstrator provides a faceted search interface to Kansa taisteli magazine articles, and links each article to further contextual data, such as related places, Wikipedia articles, troops, persons etc. based on the article metadata. Links to WarSampo demonstrators as well as further information about the project is provided at `http://www.sotasampo.fi/en/`.

WarSampo is implemented using the "7-star" Linked Data Finland platform[6] [7], based on Fuseki[7] with a Varnish Cache[8] front end for serving LOD. As a first official LOD publication, the casualty data from the National Archives is already publicly available for everyone to use[9].

[6] http://www.ldf.fi
[7] http://jena.apache.org/documentation/serving_data/
[8] https://www.varnish-cache.org
[9] http://www.ldf.fi/dataset/narc-menehtyneet1939-45

4 Related Work and Discussion

There are several projects publishing WW1 data on the web, such as Europeana Collections 1914–1918[10], 1914–1918 Online[11], WW1 Discovery[12], Out of the Trenches[13], CENDARI[14], Muninn[15], and WW1LOD [8]. War history makes a promising use case for Linked Data because war data is heterogeneous, distributed in different countries and organizations, and written in different languages [5].

Many web sites publish data about the WW2. For example, the key datasets of WarSampo have been published in Finland by our collaborators, and in other countries many more sites are online, such as the World War II Database[16] to name one. However, there are only few works on linking WW2 data, such as [2,1]. Much of the WW2 data is still confidential because people involved in the incidents or their close relatives are still alive. WarSampo contributes to related research by initiating and fostering large scale LOD publication of WW2 data, based on event-based data modeling. Our work is funded by the Ministry of Education and Culture and Finnish Cultural Foundation.

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. Association of Computing Machinery, New York (2013)
2. Collins, T., Mulholland, P., Zdrahal, Z.: Semantic browsing of digital collections. In: Proc. of the 4th International Semantic Web Conference (ISWC 2005). Springer–Verlag (2005)
3. Doerr, M.: The CIDOC CRM – an ontological approach to semantic interoperability of metadata. AI Magazine 24(3), 75–92 (2003)
4. Hyvönen, E., Viljanen, K., Tuominen, J., Seppälä, K.: Building a national semantic web ontology and ontology service infrastructure—the FinnONTO approach. In: Proc. of the 5th European Semantic Web Conference (ESWC 2008). pp. 95–109. Springer–Verlag (2008)
5. Hyvönen, E.: Publishing and using cultural heritage linked data on the semantic web. Morgan & Claypool, Palo Alto, CA, USA (2012)
6. Hyvönen, E., Lindquist, T., Törnroos, J., Mäkelä, E.: History on the semantic web as linked data – an event gazetteer and timeline for World War I. In: Proc. of CIDOC 2012 – Enriching Cultural Heritage (2012)
7. 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, Revised Selected Papers. pp. 226–230. Springer–Verlag (2014)
8. Mäkelä, E., Törnroos, J., Lindquist, T., Hyvönen, E.: World War I as Linked Open Data (2015), http://www.seco.tkk.fi/publications/submitted/makela-et-al-ww1lod.pdf, submitted for review
9. Mäkelä, E., Hyvönen, E., Ruotsalo, T.: How to deal with massively heterogeneous cultural heritage data – lessons learned in CultureSampo. Semantic Web – Interoperability, Usability, Applicability 3(1), 85–109 (2012)

[10] http://www.europeana-collections-1914-1918.eu
[11] http://www.1914-1918-online.net
[12] http://ww1.discovery.ac.uk
[13] http://www.canadiana.ca/en/pcdhn-lod/
[14] http://www.cendari.eu/research/first-world-war-studies/
[15] http://blog.muninn-project.org
[16] http://ww2db.com