WarSampo Data Service and Semantic Portal for Publishing Linked Open Data About the Second World War History

Eero Hyvönen(✉), Erkki Heino, Petri Leskinen, Esko Ikkala, Mikko Koho, Minna Tamper, Jouni Tuominen, and Eetu Mäkelä

Semantic Computing Research Group (SeCo), Aalto University, Espoo, Finland
{eero.hyvonen,erkki.heino,petri.leskinen,esko.ikkala,mikko.koho,
minna.tamper,jouni.tuominen,eetu.makela}@aalto.fi
http://seco.cs.aalto.fi/

Abstract. This paper presents the WarSampo system for publishing collections of heterogeneous, distributed data about the Second World War on the Semantic Web. WarSampo is based on harmonizing massive datasets using event-based modeling, which makes it possible to enrich datasets semantically with each others' contents. WarSampo has two components: First, a Linked Open Data (LOD) service WarSampo Data for Digital Humanities (DH) research and for creating applications related to war history. Second, a semantic WarSampo Portal has been created to test and demonstrate the usability of the data service. The WarSampo Portal allows both historians and laymen to study war history and destinies of their family members in the war from different interlinked perspectives. Published in November 2015, the WarSampo Portal had some 20,000 distinct visitors during the first three days, showing that the public has a great interest in this kind of applications.

1 Motivation: Second World War on the Semantic Web

Many websites publish information about the Second World War (WW2), the largest global tragedy in human history[1]. Such information is of great interest not only to historians but to potentially hundreds of millions of citizens globally whose relatives participated in the war actions, creating a shared trauma all over the world. However, WW2 information on the web is typically meant for human consumption only, and there are hardly any web sites that serve *machine-readable data* about the WW2 for digital humanists [3,5] and end-user applications to use. It is our belief that by making war data more accessible our understanding of the reality of the war improves, which not only advances understanding of the past but also promotes peace in the future.

The goal of this paper therefore is to (1) initiate and foster large scale LOD publication of WW2 data from distributed, heterogeneous data silos and (2) demonstrate and suggest its use in applications and research. We introduce the

[1] http://ww2db.com, http://www.world-war-2.info, Wikipedia, etc.

H. Sack et al. (Eds.): ESWC 2016, LNCS 9678, pp. 758–773, 2016.
DOI: 10.1007/978-3-319-34129-3_46

LOD service WarSampo Data[2] and the semantic WarSampo Portal[3] on top of it. WarSampo is to our best knowledge the first large scale system for serving and publishing WW2 LOD on the Semantic Web.

World war history makes a promising use case for Linked Data (LD) because war data is by nature heterogeneous, distributed in different countries and organizations, and written in different languages. WarSampo is based on the idea of creating a shared, open semantic data repository with a sustainable "business model" where everybody wins [8]: When an organization contributes to the WW2 LOD cloud with a piece of information, say a photograph, its description is automatically connected to related data, such as persons or places depicted. At the same time, the related pieces of information, provided by others, are enriched with links to the new data.

In the following, we first present the WarSampo Data service, and then the WarSampo Portal with six different application perspectives enriching each other via data linking and shared addressing practices. In conclusion, contributions of the system are summarized and related work discussed.

2 WarSampo Datasets, Conceptual Model, and Data Service

Datasets. The WarSampo Data Service contains datasets related to the Finnish Winter War 1939–1940 against the Soviet attack, the Continuation War 1941–1944, where the occupied areas of the Winter War were temporarily regained by the Finns, and the Lapland War 1944–1945, where the Finns pushed the Germans out of Lapland. The datasets in use are 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 (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 memoirs 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 ultimately annexed by the Soviet Union. Senate atlas (7) contains historical maps of Southern Finland, and the municipalities data (8) contains the Finnish municipalities that existed during the wartime. Organization cards (9), written after the war, document events of military units during the war. National Biography (10) contains over 6,300 biographies of Finnish national figures. In WarSampo the data related to 500 persons active during the war is utilized. Data about wartime events (11), persons (12), and army units (13) were collected from various war history text books. The RDF data in WarSampo contains at the moment 7,176,900 triples.

[2] Available at http://www.ldf.fi/dataset/warsa; SPARQL endpoint: http://ldf.fi/warsa/sparql.

[3] Available at http://sotasampo.fi; WarSampo is Sotasampo in Finnish.

Table 1. Central datasets of WarSampo.

#	Name	Providing organization	Size
1	Casualties of WW2	National Archives	94,700 death records
2	War diaries	National Archives	13,000 war diaries of troops
3	Photos & films	Defence Forces	160,000 photos & films
4	Kansa Taisteli magazine articles	The Assoc. for Military History in Finland & Bonnier	3,400 articles of veteran soldiers
5	Karelian places	Jyrki Tiittanen / National Land Survey	32,400 places of the annexed Karelia
6	Karelian maps	National Land Survey	47 wartime maps of Karelia
7	Senate atlas	National Archives	404 historical maps of Finland
8	Municipalities	National Archives	625 wartime municipalities
9	Organization cards	National Archives	ca 500 army units & ca 300 persons & 642 battles
10	National Biography	Finnish Literature Society	ca 500 biographies of wartime persons
11	Wartime events	War history books	1,000 events
12	Persons	War history books, Wikipedia	2,600 persons
13	Army units	War history books	3,200 army units

Conceptual Framework and Model. Since wars are essentially sequences of events, an obvious framework for representing them is event-based modeling. There are many approaches available for this, such as Event Ontology[4], LODE[5], SEM[6], and CIDOC CRM[7] [4]. CIDOC CRM was selected as a commonly used ISO standard (21127:2014). Another reason for the selection was that this conceptual framework is not limited to modeling events only, but can be used for modeling other WarSampo contents as well, such as war diaries, magazine articles, casualty records, and photos.

The core classes used in our event model is represented in Fig. 1 where namespaces `crm`, `dc`, and `skos` refer to CIDOC CRM, Dublin Core, and SKOS standards, respectively. Events are characterized by actors, places, and times that are represented by corresponding CIDOC CRM classes: Actors (crm:E39_Actor) are either persons (crm:E21_Person) or groups (crm:E74_Group). Persons are characterized by the following event types: birth, death, military rank promotion, and getting a medal of honor. Groups have subclasses of military units that may be involved in events where a unit is formed, the unit is renamed, the unit is joined with other units, and a person is joining the unit. There are currently 327,200 events in WarSampo. For Places, the Hipla.fi ontology of Karelian places and historical maps [11] is used, and for times CIDOC CRM time spans. Metadata about documentary objects, such as war diaries, magazine articles,

[4] http://motools.sourceforge.net/event/event.html.
[5] http://linkedevents.org/ontology/.
[6] http://semanticweb.cs.vu.nl/2009/11/sem/.
[7] http://cidoc-crm.org.

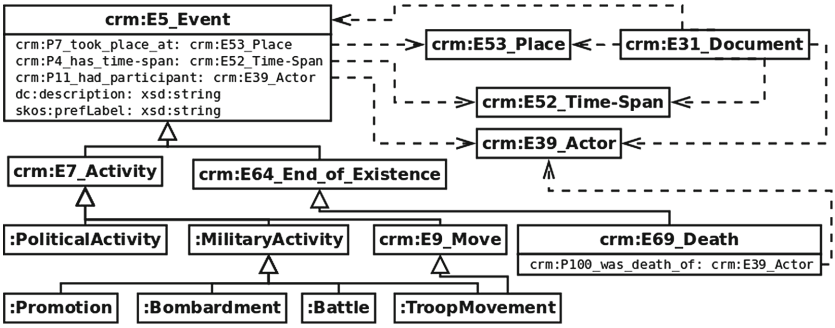

Fig. 1. Core classes of CIDOC CRM used in WarSampo.

casuality records, and photos is represented as instances of crm:E31_Document. For subject matter, the comprehensive Finnish KOKO ontology[8] of over 47,000 keyword concepts is used. Documentation about the data and metadata schemas used are available at the data service homepage[9].

Data Service. WarSampo Data is available as mutually linked open datasets. The data is provided using the "7-star" LD model [10], where the first five stars are equal to the traditional LD 5-star model [6], the 6th star is credited if the data is provided with an explicit schema, and the 7th star if the data has been validated against the schema. WarSampo was given six stars. The idea of the extra stars is to foster reuse of the data. In addition to traditional linked data services, i.e., full dataset download, URI redirection, linked data browsing, and SPARQL querying, the WarSampo Data Service provides the user with a variety of other services for data production, editing, documentation, validation, and visualization available at the hosting Linked Data Finland platform[10] [10]. The service is based on Fuseki[11] with a Varnish Cache[12] front end for serving LOD.

In contrast to the generic LOD Cloud[13], the WarSampo data cloud has a particular application domain in focus. A larger vision behind our work is that by publishing openly shared ontologies and data about WW2 for everybody to use in annotations, future interoperability problems can be prevented before they arise [7].

3 WarSampo Portal

Providing Interlinked Perspectives of War. The WarSampo Portal is not just one application, but a collection of six interlinked applications, and more are

[8] https://finto.fi/koko/en/.
[9] http://www.ldf.fi/dataset/warsa/.
[10] See http://www.ldf.fi for more details.
[11] http://jena.apache.org/documentation/serving_data/.
[12] https://www.varnish-cache.org.
[13] http://linkeddata.org.

being designed. The idea is that in order to address different end-user information needs properly, different application perspectives are needed [9,16]. For example, a first user may want to see how the war events evolve in time and geographically, a second one is interested in persons and their stories of the war, and a third one wants to do research on the casualty records of the war. The idea of providing perspectives is different from large monolithic portals like Europeana that may show only one view or search perspective of the data.

An important feature of WarSampo is that the different application perspectives can be supported without modifying the data, which would be costly given the size and complexity of the knowledge graph, but by only modifying the way the data is accessed using SPARQL. In this way new application perspectives to the data can be added more easily and independently without affecting the other perspectives.

WarSampo not only provides multiple perspectives, but also supports their interlinking using a systematic URI referencing policy. While the WarSampo Data Service is able to resolve each WarSampo URI in the traditional LD way, each application perspective is assumed to be able to resolve the URIs of its application domain as domain specific HTML pages for human usage. In a sense, each resource, e.g., a soldier in the "person" perspective, has a kind of homepage, created by the perspective, that can be linked easily to the home pages of the other perspectives, if the URI is known. Each application perspective, and also any application external to WarSampo, is able to use these ready-to-use pages via URLs. For example, an event page describing a battle event, can easily provide more information about the persons involved in the battle or the historical locations where it took place.

Many datasets in Table 1 have their own perspectives, where the user can first search data of interest and then get linked data related to them. The perspectives enrich each other via linked data. The datasets are published in the WarSampo SPARQL endpoint[14] as separate graphs. The URIs of the data resources are minted using the following template: http://ldf.fi/warsa/GRAPH/LOCAL_ID. For example, the URI http://ldf.fi/warsa/events/event_536 identifies the event "Field Marshal Mannerheim inspected the Detachment Sisu consisting of foreign volunteers in Lapua". The WarSampo Data Service documentation page contains further example URIs and SPARQL queries, e.g., one for finding events, photographs, and articles that are situated in the city of Vyborg.

The data service can be used as a basis for Rich Internet Applications (RIA). A demonstration of this is the WarSampo Portal, where *all* functionality is implemented on the client side using JavaScript, only data is fetched from the server side SPARQL endpoints. In below, the six perspectives of the WarSampo portal are presented from the point of view of end-user information needs and technological solutions.

Event-Based Perspective. The WarSampo event-based perspective[15] is aimed towards anyone interested in the course of events of the Winter and Continuation War. The events are visualized using a timeline and a map. Each event has a detailed

[14] http://ldf.fi/warsa/sparql.
[15] http://www.sotasampo.fi/events.

description and contextualizing hyperlinks to other perspectives through entities linked to the event.

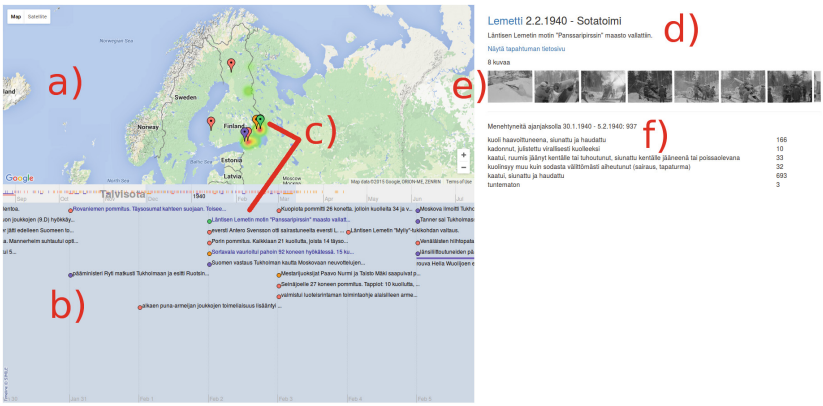

Fig. 2. Event perspective featuring a timeline and map.

Figure 2 illustrates the WarSampo event perspective. Events are displayed on a Google map (a) and on a timeline (b) that shows here events of the Winter War. When the user clicks an event, it is highlighted (c), and the historical place, time span, type, and description for the selected event are displayed (d). Photographs related to the event (e) are also shown. The photographs are linked to events based on location and time. Furthermore, information about casualties during the time span visible on the timeline is shown alongside the event description (f), and the map (a) features a heatmap layer for a visualization of these deaths.

The events can also be found and visualized through other perspectives. For example, in the Army Unit perspective, the events in which a unit participated can be viewed on maps and in time, providing a kind of graphical activity summary of the unit. In the Casualties perspective, military units of the dead soldiers are known, making it possible to sort out and visualize the personal war history of the casualties, e.g., on historical maps that come from a yet another dataset in WarSampo.

The main data sources for events were text books with event lists, including [12,13]. The pages with the lists were scanned, OCR'd, structured as CSV, and transformed into instances of CIDOC CRM event (sub)classes (cf. Fig. 1). In order to keep the visualization comprehensible, the timeline does not show minor events such as troop movements—these are visualized in the unit perspective instead (to be discussed later). The event metadata includes the description, time span, location, and participants of the event, represented using corresponding WarSampo domain ontologies.

The textual event descriptions were annotated using the ARPA automatic annotation service [15]. Automatic linking brings about the issue of name ambiguity.

Military persons mentioned in descriptions mostly have high ranks, which helps identifying them. Approaches to the place name ambiguity problem are discussed later below. Entity recognition for extracting links is still a work in progress, and conditions for it will be tweaked further to achieve a balance between precision, i.e., minimizing the amount of incorrect links, and recall, i.e., extracting as many as links as possible.

Fig. 3. Person perspective.

Person Perspective. The WarSampo person perspective application[16] is illustrated in Fig. 3. Its typical use case is someone searching for information about a relative who served in the army. On the left, the page has an input field (a) for a search by person's name. The matching names in the triple store are shown in the text field below the input. After making a selection, information about the person is shown at the top of the page (b): name, times and places of birth and death, professions, military ranks and promotions, etc. In the example case, the page shows matching photographs[17] (c), a short biography page from the National Biography[18] (d) and a set of lists linking to related events (e), military units (f), battles (g), military ranks (h), and Kansa Taisteli magazine articles (i) that mention him.

Currently the dataset consists of 96,000 persons. The data has been collected from various sources: lists of generals, lists of commanders in army corps, divisions, and regiments, lists of recipients of honorary medals like the Mannerheim

[16] http://sotasampo.fi/persons.
[17] http://sa-kuva.fi/neo?tem=webneoeng.
[18] http://www.ldf.fi/dataset/history.

Cross, casualties database, unit commanders mentioned in Organization Cards, the Finnish National Biography, Wikidata, and Wikipedia. Besides military personnel, an extract of 580 civil persons from the National Biography database and Wikidata was included in WarSampo because of their connections to WarSampo data. This set consists of persons with political or cultural significance during the wartime. The process of producing the data differed a lot depending on the used data source. For example, data lists have been scanned from a variety of documents, OCR'd, converted into CSV, and finally into RDF format. On the other hand, the casualty data of National Archives and the biographies of the National Biography had already been transformed into LOD in our earlier projects.

Some data sources, like the casualties database, provide detailed descriptions of person's life span, places, profession, marital status, etc. In contrast, sources such as the Organization Cards might only mention that, e.g., someone called *Captain Karhunen* has been in command of his unit in a certain battle. Regarding person names, we faced lots of different mentioning practices: a person might be referred to by full name (*Paavo Juho Talvela*), by initials (*P. Talvela*) or by using a combination of rank and family name (*Major General Talvela*, earlier known as *Colonel Talvela*). Recognizing whether such terms refer to the same person or not, often required extra knowledge of the person.

Person instances record only the basic properties, like family name (the only required property), forenames, a description, and provenance data, i.e., a link to the source from which the data was extracted. All other information is modeled as events, such as person's birth, death, promotion, or joining a military unit. Using the event-based approach turned out helpful especially in dealing with changing information. Consider a person's military rank: we may not know it at all, it might be a constant value during the entire wartime, or in the case of a longer military career, the rank is actually defined by a sequence of promotions. In a similar manner a person might be transferred into a different military unit and have a new commanding role in it.

The war diaries[19], data sources[20], and ranks[21] are in separate graphs. The War Diary graph has 13,043 data entries, and there are 10 data sources and 195 entries for ranks. The data includes the full range of ranks used by the Finnish Army added with some ranks used by German and Soviet Armies. Besides the military there are also some civil titles, like the ones used by the women's voluntary association *Lotta Svärd*.

Army Unit Perspective. WarSampo army unit perspective application[22] is illustrated in Fig. 4. A typical use case is someone searching for information about a specific army unit, maybe a unit where an elder relative is known to have served during the Winter War. On the left there is an input field (a) for a search by unit's name.

[19] See, e.g., http://digi.narc.fi/digi/hae_ay.ka?sartun=319.SARK.
[20] See, e.g., http://ldf.fi/warsa/actors/source3.
[21] See, e.g., http://ldf.fi/warsa/actors/ranks/Sotamies.
[22] http://sotasampo.fi/units.

The results matching unit labels in the triple store are shown in the text field below the input. The map (b) illustrates the known locations of the unit. The heatmap shows the casualties of the unit and the timeline (c) the events of the unit, e.g., dates of unit foundations, troop movements, and durations of fought battles. On the right there is a list of persons (d) known to have served in that unit. Three lists of related units are shown (e) consisting of (1) larger groups where this unit has been as a member, (2) smaller subunits being parts of this unit, and (3) otherwise related units at the same level in the hierarchy of the Finnish Army. Below this, there are additional information fields for related battles (f) and places (g), and links to entries in War Diaries (h) of the unit. There are also links to Kansa Taisteli magazine articles and photographs if they are related to the unit.

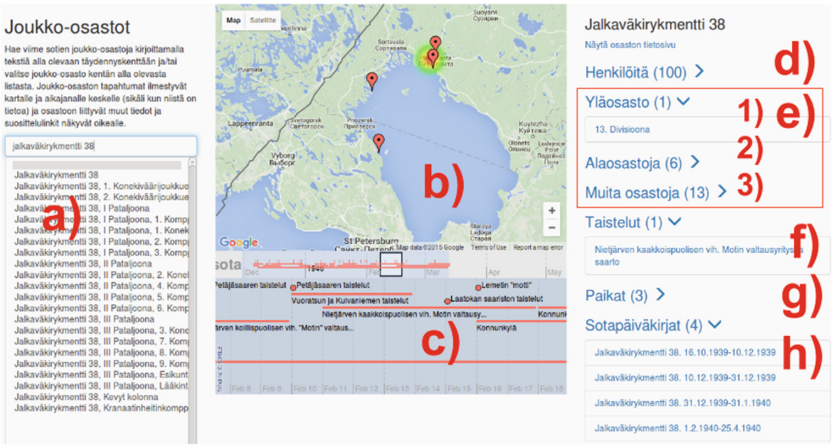

Fig. 4. Army unit perspective.

The data consists of over 3,000 Finnish army units, including Land Forces, Air Forces, Navy and its vessels, Medical Corps, stations of Anti-Aircraft Warfare and Skywatch, Finnish White Guard, and Swedish Volunteer Corps. The main sources of information have been the War Diaries and Organization Cards. The War Diaries provided an excellent starting point with about 3,000 unit labels. Currently only a part of Organization Cards are in the database, including the most important Divisions and Regiments of Infantry—during WW2 most soldiers served in Artillery and Infantry of the Land Forces, which formed the backbone of the Finnish Army.

The data in the Military Unit Ontology has been gathered simultaneously with person data. The event-based data model of a military unit is analogous to the model of a person. Also the problems regarding named entity recognition are similar in many ways. In the data sources, there are several ways of referring to a unit: by full name, e.g., *Jalkaväkirykmentti 11 (11th Infantry Regiment)*, by an abbreviation. e.g., *JR 11*, or in some cases by a nickname, e.g., *Ässärykmentti (Ace Regiment)*.

In addition, during the Winter War many units were renamed in order to confuse the enemy.

Historical Places Perspective. Most datasets used in WarSampo contain references to historical places (crm:E53_Place). If coordinates are available, places can be visualized on maps, providing a yet another perspective[23] to find and view WarSampo contents. Historical places are also essential for interlinking the datasets. For these purposes, a wartime place ontology containing place names with different levels of granularity and types (e.g., counties, municipalities, villages, bodies of water) was created as a pilot implementation of the "Finnish Ontology Service of Historical Places and Maps" [11]. After the creation of the place ontology, the other WarSampo datasets were programmatically linked to its place instances. This made it possible to build a perspective for viewing WarSampo contents on both modern and historical maps.

Figure 5 depicts the main functions of the historical places Perspective. For serendipitous browsing, all places that possess links to other WarSampo datasets can be visualized as markers or polygons on the Google map by pushing the button (a). This gives an overview of all places related to the war. In case the user is searching for a particular place, a tab for federated text search with autocompletion (b) is also provided. The search results are listed below the search field and are dynamically visualized on the map. The user can select a place by clicking on a search result row, or on a marker on the map. In the figure, the user has selected a village with the Finnish place name "Vääräkoski" that is then shown on the map with an infobox (f). By clicking the buttons (g) on the box the user can view and explore the linked events and photographs related to Vääräkoski.

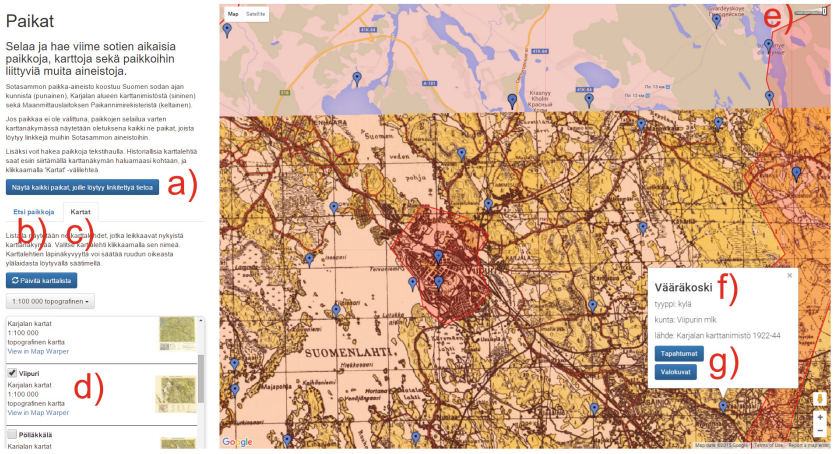

Fig. 5. Historical places perspective.

[23] http://www.sotasampo.fi/places.

In addition to the search tab described above, there is also a historical maps tab (c) on the perspective. It provides the user with a list of selectable historical maps that intersect the current Google map view. In the figure, a historical map sheet covering the city of Viipuri and its neighborhoods (d) is selected. The opacity of the historical map sheets can be adjusted with the slider (e), which allows the user to investigate both historical and modern maps at the same time, providing new insight into place names. In this case, she realizes that the place she has selected, the village "Vääräkoski" (f), can be found only from the historical map of Viipuri—obviously the village does not exist anymore.

The historical place ontology was created using four data sources: (1) a map application the National Archives of Finland (612 wartime municipalities), (2) Finnish Spatio-Temporal Ontology (polygon boundaries of the municipalities)[24], (3) a dataset of geocoded Karelian map names (35,000 map names with coordinates and place types), and (4) the current Finnish Geographic Names Registry (800,000 places). The places were modeled with a simple schema used in [11], which contains properties for the place name, coordinates, polygon, place type, and part-of relationship of the place.

The big challenge when working with place names is that place names are highly ambiguous (polysemy). There can be dozens or even hundreds of places around Finland with the same name, which presents problems for automatic annotation of description texts. Utilizing place type information is one partial solution to this problem. When linking place name mentions to the WarSampo place ontology the following order of priority was used: (1) municipality (2) town (3) village (4) body of water. House names were most ambiguous, and they were not used in automatic linking.

Another major difficulty we encountered was that different geographic data sources, such as maps used as the basis for geocoding, are overlapping, producing multiple instances of same places. A partial solution to this issue was to remove duplicate place names in advance, when two places shared a name, were close to each other, and had the same place type. However, in practice there still remained cases where it is not possible to disambiguate multiple place names without manual work.

Casualties Perspective. The casualties perspective[25] is based on the National Archives' dataset of all known Finnish casualties of WW2. The dataset consists of some 95,000 war casualty records from 1939 to 1945. The data has been originally in a relational database, which was then converted into RDF and enriched by linking it to other datasets of WarSampo. In particular, each casualty record is linked to military ranks, units, persons, and wartime municipalities. In addition, there are links to resources within the dataset, such as instances of graveyards around Finland where the deceased are buried. The casualty dataset graph consists of almost 2.5 million triples. As the dataset is large, with links to various kinds of information about each casualty, it is not straightforward to present it in an online service for users to search and browse.

[24] http://seco.cs.aalto.fi/ontologies/sapo/.
[25] http://www.sotasampo.fi/casualties.

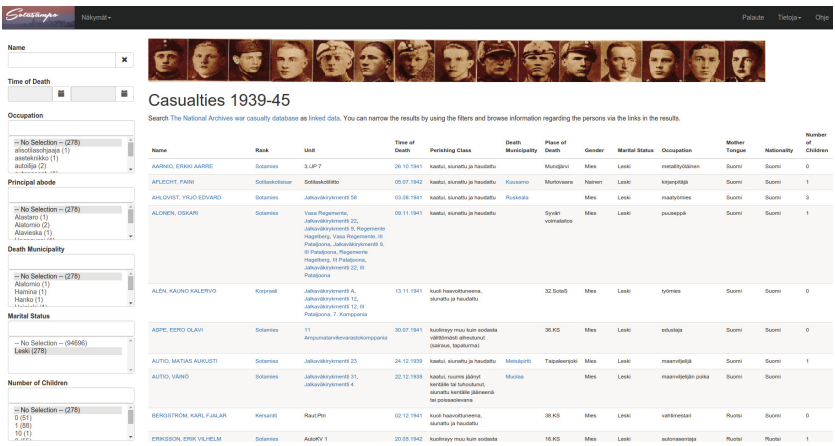

Fig. 6. Casualties perspective with one selected facet.

The casualties perspective, shown in Fig. 6, is a table-like view of the data records that can be filtered using faceted semantic search. Facets associated with the casualties are presented on the left of the interface as hierarchical facets with string search support. The number of hits on each facet category is calculated dynamically and shown to the user, so that selections leading to empty result set can be avoided. In addition, there is a special text search facet for finding persons directly by name, and a date range selector to filter the results by date of death.

In the figure, five facets are open and the other facets are not visible as they don't fit into the browser screen. The user has selected on the marital status facet the category "widow", focusing the search down to 278 killed widows of war that are presented in the table with links to further information.

Faceted search can not only be used for searching but also as a flexible tool for researching the underlying data [18]. In Fig. 6, the hit counts immediately show distributions of the killed widows along the facet categories. For example, the facet "Number of children" shows that one of the deceased had 10 children and most often (in 88 cases) widows had one child. If we next select category "one child" on its facet, we can see that two of the deceased are women and 86 are men in the gender facet.

Our faceted search engine is based purely on SPARQL queries and client side data processing in JavaScript. The system works well even with the large datasets of WarSampo, as pagination is used to limit the amount of results that are queried and displayed to the user.

The casualty records were modeled using the class crm:E31_Document with a distinct property for each facet. The property values are annotation resources selected from the corresponding ontologies, such as places. Record instances refer also to events, e.g., the death events of persons.

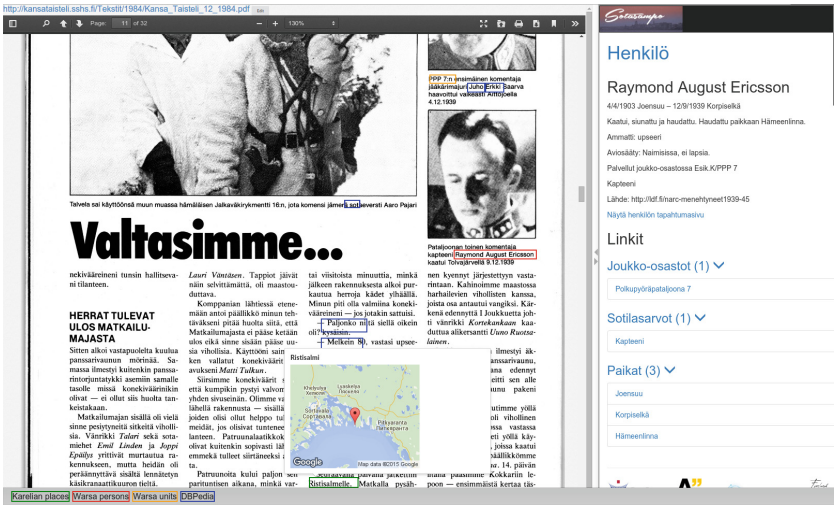

Fig. 7. The Contextual Reader interface targeting the Kansa Taisteli magazine articles.

Magazine Article Perspective. This application[26] is for searching and browsing textual articles relating to WW2. Here, the content are the 3,357 Kansa Taisteli magazine articles published by Sotamuisto in 1957–1986, containing mostly memoirs of soldiers related to WW2. The purpose of the perspective is two-fold: (1) to help a user find Kansa Taisteli articles of interest using faceted semantic search and, (2) to provide context to the found articles by extracting links to related WarSampo data from the texts.

The start page of the magazine article perspective is a faceted search browser similar to the one in the casualties perspective (cf. Fig. 6). Here, the facets allow the user to find articles by filtering them based on author, issue, year, related place, army unit, or keyword. Some of the underlying properties, such as the year and issue number of the magazines, are hierarchical and represented using SKOS. The hierarchy is visualized in the appropriate facet, and can be used for query expansion: by selecting an upper category in the facet hierarchy one can perform a search using all subcategories.

After the user has found an article of interest, she can click on it, and the digitized article appears on the screen in the CORE Contextual Reader interface [17]. Depicted in Fig. 7, CORE is able to automatically and in real time annotate PDF and HTML documents with recognized keywords and named entities, such as army units, places, and person names. These are then encircled with colored boxes indicating the linked data source. By hovering the mouse over a box, linked data from the data source is shown to the user, providing contextual information for an enhanced reading experience. In Fig. 7 the user is hovering on the identified place *Ristisalmi*,

[26] http://www.sotasampo.fi/articles.

which is then shown on a map for contextualization. If further contextual information is desired, the user can click on an entity to open the WarSampo page for that entity on a pane to the right of the reader interface. In Fig. 7, for example, detailed data are shown about *Raymond August Ericsson*, one of the battalion commanders discussed in the article.

The Kansa Taisteli magazine articles used in the interface have been manually scanned into PDF format by a member of the Association for Military History in Finland, Timo Hakala, and made available on the association's web site[27] in collaboration with the current copyright holder, Bonnier Publications. Our search application additionally makes use of a separate CSV file containing metadata for the 3,357 articles, also manually crafted by Timo Hakala.

After transforming the metadata into instances of documents (crm: E31_Document) and linking it with the WarSampo domain ontologies, the article dataset was further enriched with subject matter keywords by using the ARPA automatic text annotation service in the same way as with the other datasets. The extracted keywords were resources indicating military units, military persons, and places mentioned in the article text. These resources are used as the basis for the keyword facet in searching. The enriched metadata of the articles contains approximately 44,000 triples in total. The metadata is based on Dublin Core, where in addition to some standard properties like *dc:title*, there are object properties corresponding to each search facet, which facilitate the search.

A challenge faced during the linking and annotating of the Kansa Taisteli articles was the quality of the data. For example, because the magazines were manually scanned in a laborious process, full-page advertisements were sometimes not included. However, when locating the articles inside the PDFs based on the metadata, this threw off the reader sometimes even by multiple pages. A more serious concern was errors of the OCR process that caused challenges for the automatic annotation process. For example, unit names as abbreviations are inflected in Finnish by appending a *:* and the inflection ending. However, in OCR, character *:* was often read as *i* or *z*. Luckily, being a specialized domain with rigid conventions for writing, e.g., units and ranks, most of these errors could be corrected using a host of 135 regular expression rules.

This still left the problem of semantic disambiguation; in this case this concerned named entity recognition of persons, places, and military units. Formal evaluation on the automatic annotation process has not been made, but based on an informal evaluation, the final outcome is useful for its purpose even if the annotations are incomplete and some errors remain.

4 Related Work, Discussion, and Future Work

There are several projects publishing linked data about the World War I on the web, such as Europeana Collections 1914–1918[28], 1914–1918 Online[29], WW1

[27] http://kansataisteli.sshs.fi.
[28] http://www.europeana-collections-1914-1918.eu.
[29] http://www.1914-1918-online.net.

Discovery[30], Out of the Trenches[31], CENDARI[32], Muninn[33], and WW1LOD [14]. There are few works that use the Linked Data approach to WW2, such as [1,2] and Open Memory Project[34] on holocaust victims.

Our results suggest that large heterogeneous datasets of war history can be interlinked with each other through events in ways that provide insightful multiple perspectives for the historians and laymen to the data. Given the wide, deep, and sentimental interest in war history among the public and researchers, we envision that war history will become an important domain for Linked Data applications.

We have also learned that even in the rural northern parts of Europe, massive amounts of WW2 data can be found and opened for public use. We have initially dealt with less than 100,000 people involved in war events. However, there is also data available about hundreds of thousands of soldiers who survived the war only in Finland. Managing the data, and providing it for different user groups, suggests serious challenges when dealing with, e.g., the war events in the central parts of Europe, where the amount of data is orders of magnitude larger than in Finland, multilingual, and distributed in different countries. For example, solving entity resolution problems regarding historical place names and person names can be difficult. However, it seems that Linked Data is a promising way to tackle these challenges.

Future work on WarSampo includes, e.g., end user evaluations, where the portal is compared with existing legacy database services in searching for WW2 materials, and where the usability of the portal is tested in its use cases. We also plan to continue our work on automatic annotation of texts.

Acknowledgements. Jérémie Dutruit created the first RDF version of the casualties data, Jyrki Tiittanen geocoded the Karelian places dataset, Hanna Hyvönen rectified the historical maps on modern ones, Timo Hakala provided the Kansa Taisteli CSV metadata, and Kasper Apajalahti transformed it into RDF. Our work is funded by the Ministry of Education and Culture and Finnish Cultural Foundation. Wikidata Finland project financed rectifying of the historical maps.

References

1. de Boer, V., van Doornik, J., Buitinck, L., Marx, M., Veken, T.: Linking the kingdom: enriched access to a historiographical text. In: Proceedings of the 7th International Conference on Knowledge Capture (KCAP 2013), pp. 17–24. ACM, June 2013
2. Collins, T., Mulholland, P., Zdrahal, Z.: Semantic browsing of digital collections. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 127–141. Springer, Heidelberg (2005)

[30] http://ww1.discovery.ac.uk.
[31] http://www.canadiana.ca/en/pcdhn-lod/.
[32] http://www.cendari.eu/research/first-world-war-studies/.
[33] http://blog.muninn-project.org.
[34] http://www.bygle.net/wp-content/uploads/2015/04/Open-Memory-Project_3-1.pdf.

3. Crymble, A., Gibbs, F., Hegel, A., McDaniel, C., Milligan, I., Posner, M., Turkel, W.J. (eds.): The Programming Historian, 2nd edn. (2015). http://programminghistorian.org/
4. Doerr, M.: The CIDOC CRM - an ontological approach to semantic interoperability of metadata. AI Mag. **24**(3), 75–92 (2003)
5. Graham, S., Milligan, I., Weingart, S.: Exploring Big Historical Data: The Historian's Macroscope. Imperial College Press, London (2015)
6. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, 1st edn. Morgan & Claypool, Palo Alto (2011). http://linkeddatabook.com/editions/1.0/
7. Hyvönen, E.: Preventing interoperability problems instead of solving them. Semantic. Web J. **1**(1–2), 33–37 (2010)
8. Hyvönen, E.: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, Palo Alto (2012)
9. 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: Proceedings of CIDOC 2012 - Enriching Cultural Heritage, CIDOC, June 2012
10. 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: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) ESWC Satellite Events 2014. LNCS, vol. 8798, pp. 226–230. Springer, Heidelberg (2014)
11. Hyvönen, E., Tuominen, J., Ikkala, E., Mäkelä, E.: Ontology services based on crowdsourcing: case national gazetteer of historical places. In: Proceedings of 14th International Semantic Web Conference (ISWC 2015), Posters and Demonstrations Track. CEUR Workshop Proceedings, vol. 1486, October 2015
12. Leskinen, J., Juutilainen, A. (eds.): Jatkosodan Pikkujättiläinen. WSOY, Finland (2005)
13. Leskinen, J., Juutilainen, A. (eds.): Talvisodan pikkujättiläinen, 4th edn. WSOY, Finland (2006)
14. Mäkelä, E., Törnroos, J., Lindquist, T., Hyvönen, E.: World War 1 as Linked Open Data (2015), submitted for review. http://seco.cs.aalto.fi/publications/
15. Mäkelä, E.: Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) ESWC Satellite Events 2014. LNCS, vol. 8798, pp. 424–428. Springer, Heidelberg (2014)
16. 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)
17. Mäkelä, E., Lindquist, T., Hyvönen, E.: CORE - a contextual reader based on linked data. In: Proceedings of Digital Humanities 2016, long papers, July 2016
18. Tunkelang, D.: Faceted Search. Retrieval, and Services, Morgan & Claypool, Palo Alto, CA, USA, Synthesis Lectures on Information Concepts (2009)