Linked Open Data and Ontology Infrastructure for Second World War History

Mikko Koho[1], Erkki Heino[1,2], Petri Leskinen[1], Minna Tamper[1,2], Esko Ikkala[1],
Eetu Mäkelä[1,2], Jouni Tuominen[1,2], and Eero Hyvönen[1,2]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland and
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
`http://seco.cs.aalto.fi, http://heldig.fi`
`firstname.lastname@aalto.fi`

Abstract. Data about the Second World War (WW2) is heterogeneous and distributed in different organizations and countries. This paper argues that in order to create aggregated global views of the war, a shared semantic infrastructure is needed, including data models of the real world events, metadata schemas for presenting their documentation, a data harmonization model for data aggregation, and shared domain ontologies for populating the schemas in an interoperable way. As a solution, a Linked Open Data service is presented for publishing data about Finland in WW2. The service is based on W3C Semantic Web standards and best practices, including content negotiation, SPARQL API, download, automatic documentation, and other services supporting re-use of the data. The ontologies and data in the service, totalling ca. 9 million triples, is in use in seven end-user applications of the WarSampo portal, that have had tens of thousands of end-users.

1 Introduction

Plenty of data about WW2 exists in different organizations in a multitude of both unstructured and structured formats. Gathering and unifying the data of the whole war is not a simple task. This is nonetheless needed in order to create a global view of the war, and to attain a deeper understanding of its history.

WarSampo collects data related to the Finnish wars in WW2 [8], and publishes this data as Linked Data [1] on an open SPARQL endpoint, and user-friendly online interfaces in the WarSampo portal[3], which has had tens of thousands of users. WarSampo is to our best knowledge the first large scale system for serving and publishing WW2 LOD on the Semantic Web.

Our previous papers on WarSampo have presented the vision and overview of the system especially from the use case and end-user application perspectives [8]. In [5,12] data creation was concerned from the Named Entity Linking point of view. In contrast, this resource description paper presents the LOD infrastructure used in WarSampo in detail. In the following, we first present the event-based data model and datasets of the service, and discuss then the data service and related works.

[3] `http://sotasampo.fi`

2 Event-Based Model

Since wars are essentially sequences of events, an obvious choice for representing war history is event-based modeling. There are many approaches for modeling events [11]. We use the CIDOC Conceptual Reference Model (CRM)[4] [4] as it is an ISO standard, and it can be used for modeling other war history content also, such as war diaries, magazine articles, casualty records, and photos. Using events also makes it possible to describe changes of the status of different entities, such as people and military units. Furthermore, using a common model for all the datasets makes querying data more uniform.

Details of the CRM classes used in WarSampo is given in Table 1, which shows the CRM class, number of instances of that class, and what is it used for. Some resources are instances of multiple CRM classes. Names of resources are annotated with `skos:prefLabel` and `skos:altLabel`, while information sources are given with `dct:source`, and textual descriptions with `dct:description`.

Table 1. The CIDOC CRM classes used in WarSampo.

Class	Nr. of instances	Used for
E4_Period	5	Conflicts
E5_Event	554,605	War time events
E7_Activity	361,349	Military activity
E9_Move	684	Events
E21_Person	99,703	Actors
E24_Physical_Man-Made_Thing	200	Medals
E31_Document	288,350	Photographs, death records, war diaries, data sources
E38_Image	163,894	Photographs
E39_Actor	115,629	Actors
E52_Time-Span	34,768	Time span of an event
E53_Place	33,043	Historical places
E55_Type	388	Military ranks, unit types
E65_Creation	114,537	Photography
E66_Formation	16,739	Events
E67_Birth	96,755	Events
E68_Dissolution	7,701	Events
E69_Death	96,345	Events
E74_Group	15,926	Actors

Most resource URIs are of the form `http://ldf.fi/warsa/`*`dataset/id`* where *dataset* is a name of a dataset, and *id* is an identifier consisting of a prefix and a running number. For example:

`http://ldf.fi/warsa/photographs/sakuva_57717`.

Based on the URIs, instances of central resources types such as persons, units, places have HTML homepages, whose URL is

[4] `http://cidoc-crm.org`

`http://www.sotasampo.fi/en/page?uri=`*uri*.

In this way, links to related homepages can be created easily across different application perspectives. The URIs for domain ontologies are contained within the dataset namespace. The data in the data service is separated into multiple graphs, so that each dataset is contained in it's own graph.

The downside of using an event-based model for all the datasets is its complexity: photographs are, for example, modeled as an image and an event creating it. This can lead to complex queries and presenting the data becomes more complicated. In such cases, CRM was used for data harmonization and aggregation, but we also kept simple Dublin Core like metadata in the repository.

The WarSampo adheres to the Linked Data 5-star publishing principles. More over, there is an online documentation[5] available, to extend the LD publication quality to the sixth star of the 7-star model [7].

3 WarSampo Linked Open Data Cloud

The WarSampo datasets and linkage between them is presented in Fig. 1, which also shows the number of links and the direction of linking. The main datasets are discussed in greater detail below, with the graph names, amounts of instances and triples, and main classes and properties. Namespaces are omitted.

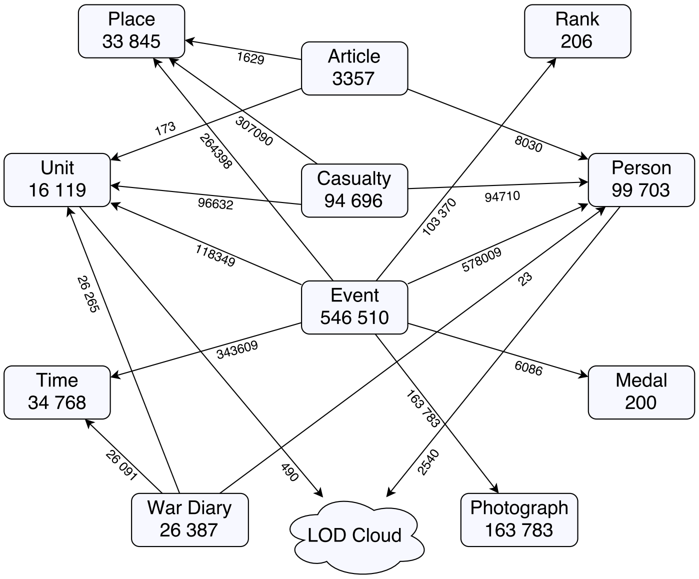

Fig. 1. Linkage between the main WarSampo datasets.

[5] `http://www.essepuntato.it/lode/owlapi/http://ldf.fi/schema/warsa`

War diaries

Graph:	`http://ldf.fi/warsa/events/diaries`
Contains:	26 387 instances, 131 657 triples
Core classes:	`crm:E31_Document`
Properties:	`crm:P70_documents`, `dc:hasFormat`, `crm:P4_has_time-span`
Source:	National Archives

Each resource in the war diaries graph contains metadata of a War Diary, e.g., `dct:hasFormat` links to corresponding webpage at the National Archives[6], `crm:P4_has_time-span` indicates when the diary was written, and a `crm:P70_documents` links to corresponding military unit or person.

Wartime Events

Graph:	`http://ldf.fi/warsa/events`
Contains:	1922 instances, 19 269 triples
Core classes:	`crm:E5_Event`, `crm:E69_Death`
Properties:	`skos:related`, `crm:P100_was_death_of`, `crm:P11_had_participant`, `crm:P14_carried_out_by`, `crm:P4_has_time-span`, `crm:P7_took_place_at`, various custom properties
Source:	War history books, Wikipedia, Finnish Defense Forces

The events graph contains events extracted from various sources. The events provide an overview of the war focusing on Finland. Events can be found in other graphs in WarSampo as well, but all notable war events are in this graph.

Each event's type is `crm:E5_Event` or its subclass. Each event has a textual representation (`skos:prefLabel`, `dct:description`), and a time-span, and information where the event occurred, if applicable, linking the event to place ontologies. The events are linked to actors by several properties. The links to places and actors have mostly been generated automatically [5].

Wartime Actors

Graph:	`http://ldf.fi/warsa/actors`
Contains:	554 141 instances, 3 248 389 triples
Core classes:	`crm:E39_Actor`, `crm:E21_Person`, `crm:E67_Birth`, `crm:E69_Death`, `crm:E68_Dissolution`
Properties:	`foaf:familyName`, `foaf:firstName`, `owl:sameAs`, custom properties
Source:	War history books, Finnish Defence Forces, National Archives, Wikipedia

Each actor's type is a subclass of `crm:E39_Actor`. There are two kind of actors: people (`crm:E21_Person`, `:MilitaryPerson`), and groups (`:MilitaryGroup`). This graph also contains the events relating to individual actors, e.g. person's birth and death, promotion, joining an unit, wounding, disappearing or getting awarded with a medal, or unit's foundation, movement, renaming, joining under another unit or dissolution. The actor information is collected from various sources: scanned books and cards, datasheets, and related databases.

[6] see e.g. `http://digi.narc.fi/digi/slistaus.ka?ay=37910`

Times

Graph:	`http://ldf.fi/warsa/events/times`
Contains:	34 768 instances, 139 072 triples
Core classes:	`crm:E52_Time-Span`
Properties:	`crm:P82a_begin_of_the_begin`, `crm:P82b_end_of_the_end`

The times graph contains all the time-spans used throughout the WarSampo. Each time-span is an instance of the CRM time-span class `crm:E52_Time-Span`. As it is not generally known e.g. at what time during a day an event has occurred, the CRM property *P82 at some time within* is used.

Historical Places and Maps

Graph:	`http://ldf.fi/warsa/places/karelian_places`
Contains:	32 403 instances, 159 898 triples
Core classes:	`crm:E53_Place`
Properties:	`wgs84:lat`, `wgs84:long`, `rdf:type`, `gs:sfWithin`
Source:	Jyrki Tiittanen

Graph:	`http://ldf.fi/warsa/places/municipalities`
Contains:	625 instances, 5895 triples
Core classes:	`crm:E53_Place`
Properties:	`schema:polygon`, `wgs84:lat`, `wgs84:long`, `gs:sfWithin`
Source:	National Archives of Finland

As most WarSampo datasets contain references to place names during WW2 in Finland, a Finnish WW2 place ontology was created. Its data comes from four sources: 1) The National Archives' wartime municipalities, 2) the Finnish Spatio-Temporal Ontology describing the regions of the Finnish municipalities in different times[7], 3) a dataset of geocoded Karelian map names, and 4) the current Finnish Geographic Names Registry (PNR). In addition, some 450 historical map sheets from two atlases were rectified on modern maps.

The Finnish WW2 place ontology is a snapshot of places and regions with different levels of granularity and types (e.g. counties, municipalities, towns, villages, bodies of water) in Finland during the war years 1939–1945. The places were modeled with a simple schema [6], which contains properties for the place name, coordinates, polygon, place type, and part-of relationship of the place. Because the place ontology covers only a limited time period, there is no need to model the temporal changes of places and regions (which are usually not available) within this time period.

Wartime Photographs

Graph:	`http://ldf.fi/warsa/photographs`
Contains:	278 210 instances, 2 817 997 triples
Core classes:	`crm:E31_Document`, `crm:E38_Image`
Properties:	various properties
Source:	Photograph Archive of the Finnish Defense Forces (SA-kuva)

[7] http://seco.cs.aalto.fi/ontologies/sapo/

The photographs graph contains the metadata of over 160 000 historical photographs taken by Finnish soldiers during WW2. The `:Photograph` class is a subclass of both `crm:E31_Document` and `crm:E38_Image`, and represents a photograph. There is also a photography event class (`:Photography`), which is a subclass of `crm:E65_creation` and represents the taking (i.e., creation) of photographs, and is used for harmonization with the CIDOC CRM. The amount of photography events has been reduced so that some photographs share the same event – i.e., photographs that have been taken the same day and have the same description. Modeling the photographs using events has the benefit of making it possible to handle them the same way as other event-based entities, e.g., placing them on a timeline.

Magazine Articles

Graph:	`http://ldf.fi/warsa/articles`
Contains:	3357 instances , 72,147 triples
Core classes:	`crm:E31_Document`, `skos:Concept`
Properties:	`dc:title`, `dc:hasFormat`, `crm:P3_has_note`, `crm:P67_refer_to`, `skos:prefLabel`, various custom properties
Source:	The Association for Military History in Finland

The graph consists of Kansa Taisteli magazine articles with metadata, which have been automatically linked to ontologies from the magazine article texts [12,5]. Used ontologies are WarSampo ontologies, such as actors and places, and external ontologies, such as the Finnish DBpedia and the KOKO ontology[8].

The articles are instances of the document class `:Article` which is a subclass of `crm:E31:Document`. The articles are described bibliographically (e.g., title, author, page, issue, volume), content-wise (e.g., place, event, unit, mentionsPlace, mentionsUnit, mentionsPerson), and contextualizing or technical data (e.g., hasFormat, order). The main principle for modeling was usability in a faceted search application.

Casualties during the Finnish wars 1939–1945

Graph:	`http://ldf.fi/narc-menehtyneet1939-45/`
Contains:	95,375 instances, 2,350,186 triples
Core classes:	`crm:E31_Document`, `skos:Concept`
Properties:	`crm:P70_documents`, `skos:prefLabel`, various custom properties [9]
Source:	National Archives

The dataset of the Finnish war casualties from the Finnish National Archives consists of about 95,000 death records. It has served as the primary source of person instances in WarSampo. The death records are linked to WarSampo person instances, military units, military ranks, and wartime municipalities [9].

The death records are instances of the document class `crm:E31_Document`, and are described with 31 properties based on the original dataset. These properties convey information about, e.g., the deceased person's occupation, number

[8] `https://finto.fi/koko/en/`

of children, marital status, and burial place. We used a simple tabular-like data model for the death records, is to keep the data model simple and number of triples low, which allows quicker response times for SPARQL queries from real-time applications, such as the casualties perspective of WarSampo.

From each death record, there is a `crm:P70_documents` relation to the corresponding WarSampo person instance.

4 Discussion and Related Work

The WarSampo LOD service is available[9] on the Linked Data Finland platform [7]. The URIs used are resolvable to both human and computer users. The data and ontologies are available via open SPARQL endpoints[10], with the open Creative Commons 4.0 license, and can be downloaded, too. To support data re-use, the service also provides additional information about the dataset, like automatically generated schema documentation and example SPARQL queries. Our goal is not only to support the seven application perspectives of the WarSampo semantic portal [8], but to lay a foundation for a general infrastructure for creating interoperable Linked Data about the WW2 in Finland and beyond. Indeed, two new projects for creating linked data about ca. 4500 prisoners of war and military grave yards (over 3000 photos) are underway, and the data has been re-used by a commercial company in a communal portal[11].

There are few works that use the Linked Data approach to WW2, such as [3,2]. The Open Memory Project[12] publishes Linked Data about WW2 holocaust victims, using the Event Ontology[13] to model some events. Several projects have created linked data about the World War I, including Europeana Collections 1914–1918[14], 1914–1918 Online[15], WW1 Discovery[16], Out of the Trenches[17], and Muninn[18]. In contrast to these, WarSampo data service employs an event-based approach extending CIDOC CRM, continuing here our own earlier work on modeling WW1 [10], with an emphasis of creating shared domain ontologies and linked data services for everybody to use.

Acknowledgements Our work is funded by the Open Science and Research Initiative[19] of the Finnish Ministry of Education and Culture, the Finnish Cultural Foundation, and the Academy of Finland.

[9] `http://www.ldf.fi/dataset/warsa`

[10] `http://ldf.fi/warsa/sparql`

[11] `http://www.sotapolku.fi`

[12] `http://www.bygle.net/wp-content/uploads/2015/04/Open-Memory-Project_3-1.pdf`

[13] `http://motools.sourceforge.net/event/event.html`

[14] `http://www.europeana-collections-1914-1918.eu`

[15] `http://www.1914-1918-online.net`

[16] `http://ww1.discovery.ac.uk`

[17] `http://www.canadiana.ca/en/pcdhn-lod/`

[18] `http://blog.muninn-project.org`

[19] `http://openscience.fi/`

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data–The Story So Far. Semantic services, interoperability and web applications: emerging concepts pp. 205–227 (2009)
2. 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)
3. 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 (November 2005)
4. Doerr, M.: The CIDOC CRM – an ontological approach to semantic interoperability of metadata. AI Magazine 24(3), 75–92 (2003)
5. Heino, E., Tamper, M., Mäkelä, E., Leskinen, P., Ikkala, E., Tuominen, J., Koho, M., Hyvönen, E.: Named entity linking in a complex domain: Case second world war history. In: Language, Technology and Knowledge. Springer (2017), in press
6. Hyvönen, E., Ikkala, E., Tuominen, J.: Linked data brokering service for historical places and maps. In: Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe). pp. 39–52. CEUR Workshop Proceedings (May 2016), `http://ceur-ws.org/Vol-1608/#paper-06`, vol 1608
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 (May 2014)
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 (2016)
9. Koho, M., Hyvönen, E., Heino, E., Tuominen, J., Leskinen, P., Mäkelä, E.: Linked Death - Representing, Publishing, and Using Second World War Death Records as Linked Open Data. In: The Semantic Web: ESWC 2016 Satellite Events. Springer (June 2016)
10. 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 (2016), in press.
11. Rovera, M.: A knowledge-based framework for events representation and reuse from historical archives. In: Proceedings of the 13th International Conference on The Semantic Web. Latest Advances and New Domains-Volume 9678. pp. 845–852. Springer (2016)
12. Tamper, M., Leskinen, P., Ikkala, E., Oksanen, A., Mäkelä, E., Heino, E., Tuominen, J., Koho, M., Hyvönen, E.: AATOS: a Configurable Tool for Automatic Annotation. In: Language, Technology and Knowledge 2017. Springer (2017), in press.