WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data

Mikko Koho [a,*], Esko Ikkala [a], Petri Leskinen [a], Minna Tamper [a], Jouni Tuominen [a,b], and Eero Hyvönen [a,b]

[a] *Semantic Computing Research Group (SeCo), Aalto University, Department of Computer Science, Finland*
E-mail: firstname.lastname@aalto.fi
[b] *HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland*
E-mail: firstname.lastname@helsinki.fi

Abstract. The Second World War (WW2) is arguably the most devastating catastrophe of human history, a topic of great interest to not only researchers but the general public. However, data about the Second World War is heterogeneous and distributed in various organizations and countries making it hard to utilize. In order to create aggregated global views of the war, a shared ontology and data infrastructure is needed to harmonize information in various data silos. This makes it possible to share data between publishers and application developers, to support data analysis in Digital Humanities research, and to develop data-driven intelligent applications. As a first step towards these goals, this article presents the WarSampo knowledge graph (KG), a shared semantic infrastructure, and a Linked Open Data (LOD) service for publishing data about WW2, with a focus on Finnish military history. The shared semantic infrastructure is based on the idea of representing war as a spatio-temporal sequence of events that soldiers, military units, and other actors participate in. The used metadata schema is an extension of CIDOC CRM, supplemented by various military historical domain ontologies. With an infrastructure containing shared ontologies, maintaining the interlinked data brings upon new challenges, as one change in an ontology can propagate across several datasets that use it. To support sustainability, a repeatable automatic data transformation and linking pipeline has been created for rebuilding the whole WarSampo KG from the individual source datasets. The WarSampo KG is hosted on a data service based on W3C Semantic Web standards and best practices, including content negotiation, SPARQL API, download, automatic documentation, and other services supporting the reuse of the data. The WarSampo KG, a part of the international LOD Cloud and totalling ca. 14 million triples, is in use in nine end-user application views of the WarSampo portal, which has had over 400 000 end users since its opening in 2015.

Keywords: Linked Open Data, Semantic Web, Military History, World War II, Finland, Cultural Heritage, Digital Humanities

1. WarSampo Initiative

Plenty of information about WW2 is published every year in books, articles, news, web sites and services, documentaries, and films for humans to consume. This information is scattered in various military, governmental, cultural heritage, and other organizations, making it hard to find and use. Furthermore, the information is seldom published as data for reuse in computational analyses and applications. Gathering, extracting, and harmonizing information about the war is needed in order to create comprehensive global views of the war and history but this is not a simple task. This applies also to microhistory: for example, finding out the details of what happened to a perished relative during the war can be quite tedious, involving studying and aggregating data about him/her from several registries and data sources. Without harmonized, clean data, the data analysis of large military historical datasets, such as death records, would be difficult in Digital Humanities Research [1, 2]. Combining information from various sources facilitates answering the complex societal research questions of "new military history" scholars [3].

*Corresponding author. E-mail: firstname.lastname@aalto.fi.

The goal of the *WarSampo – Finnish Second World War on the Semantic Web* initiative[1] is to study and show how Linked Data [4] (LD) can help in solving tasks like these [5]. The initiative collects military historical data related to Finland in the Second World War (WW2). The data is published as Linked Open Data (LOD) in an open SPARQL endpoint on top of which the WarSampo portal[2] has been created, including nine application perspectives to the data. The portal, targeted to both researchers and the public at large, has had 550 000 end users since its opening in 2015. The WarSampo data service and portal were awarded with the LODLAM Challenge Open Data Prize in 2017 in Venice. The data forms an integrated interlinked 5-star LOD publication, and is part of the global LOD Cloud[3].

The WarSampo *knowledge graph* (KG) was published initially in 2015. The KG was first used by seven different application perspectives in the WarSampo portal, via only the SPARQL API [5]. The idea was to show that anyone could easily use the data dynamically on the client side. In 2017, by the centennial of Finnish independence, a new eighth application perspective of war cemetery data and related photographs[4] was released [6], a further demonstration of this idea. Finally, in 2019, a ninth application based on yet another dataset of ca. 5000 prisoners of war was aligned with the WarSampo KG and will be released [7] in November 2019.

This dataset description complements our other publications about WarSampo by presenting in detail the KG, including the process of maintaining the data.

2. Related Work

The problem of combining and using heterogeneous cultural heritage datasets is a common problem in using Linked Data for Digital Humanities [8, 9] and in Digital History [10]. Historical knowledge contextualization and visualization with experiences from the VICODI project are represented in [11], which also discusses general problems faced when modelling history with ontologies. Several humanities and cultural heritage related projects have used the *CIDOC Conceptual Reference Model (CRM)* [12][5].

Several projects have published linked data about the World War I on the web, such as Europeana Collections 1914–1918[6], 1914–1918 Online[7], WW1 Discovery[8], CENDARI[9], Muninn[10], and WW1LOD [13]. There are also a few works that have used the Linked Data approach to WW2, such as [14–16] and a LOD system on WW2 holocaust victims [17].

Our own previous research on WarSampo first presented the vision and overview of the system especially from the use case and end-user application perspectives [5, 18]. In [19] data integration was concerned from the named entity linking (NEL) point of view. The maintenance problem of the interlinked dataset has been explored in [20]. Work on creating and using individual parts of the KG has been presented in several previous publications [6, 7, 21–24].

This article is organized as follows. The next Section presents the source datasets. Section 4 discusses how the information in the source datasets was harmonized and presents the event-based data model. The data transformation process is presented in Section 5. An analysis of the data quality is given in Section 6. The stability and usefulness of the data are discussed in Sections 7 and 8, respectively, conclusion in Section 9.

3. Source Datasets

Table 1 lists the heterogenous source datasets of WarSampo. The data comes from several Finnish organizations, such as the National Archives of Finland, the Finnish Defence Forces, and the National Land Survey of Finland. Some source datasets have been created as part of the WarSampo project and related research.

The core dataset of the system is the casualty database (source number 1 in Table 1) of the National Archives that contains detailed information about virtually every person killed in action in Finland during the WW2. A key goal of WarSampo is to reassemble the life stories of the soldiers by gathering information about them via data linking. For this purpose, data about the military units (5) and their history (6), in-

[1] The initiative and publications are presented in the initiative homepage: https://seco.cs.aalto.fi/projects/sotasampo/en/.
[2] http://sotasampo.fi/en
[3] http://linkeddata.org
[4] https://seco.cs.aalto.fi/projects/sotasampo/hautausmaat/
[5] A list of CIDOC CRM use cases can be found at: http://www.cidoc-crm.org/useCasesPage.
[6] http://www.europeana-collections-1914-1918.eu
[7] http://www.1914-1918-online.net
[8] http://ww1.discovery.ac.uk
[9] http://www.cendari.eu/research/first-world-war-studies/
[10] http://blog.muninn-project.org

Table 1
Source datasets of WarSampo, grouped by providing organization. Numbers in the article are rounded to 3 significant digits.

#	Source Dataset	Providing Organization	Used Content	Source Format
1	Casualties of WW2	The National Archives of Finland	94 700 person records	spreadsheet
2	War diaries	The National Archives of Finland	26 400 war diaries with metadata, 9850 units, and 12 people	spreadsheet
3	Senate atlas	The National Archives of Finland	414 historical maps of Finland	digital images
4	Municipalities	The National Archives of Finland	625 wartime municipalities	digital text
5	Organization cards	The National Archives of Finland	132 military units & 279 people & 642 battles	digital images, PDF documents
6	Units of The Finnish Army 1941–1945	The National Archives of Finland	8810 military units	digital text, PDF document
7	Wartime photographs	The Finnish Defence Forces	164 000 photos with metadata, 1740 people	spreadsheet, API access
8	Kansa Taisteli magazine articles	The Association for Military History in Finland, Bonnier Publications	3360 articles by war veterans	spreadsheet, PDF documents
9	Karelian places	The National Land Survey of Finland	32 400 places of the annexed Karelia	spreadsheet
10	Karelian maps	The National Land Survey of Finland	47 wartime maps of Karelia	digital images
11	Finnish Place Name Register	The National Land Survey of Finland	798 000 contemporary place names	XML
12	National Biography	The Finnish Literature Society	699 biographies	spreadsheet
13	War cemeteries	The Central Organization of Finnish Camera Clubs	672 cemeteries & 2450 photographs	spreadsheet, digital images
14	Prisoners of war	The National Prisoners of War Project	4450 person records	spreadsheet
15	Wikipedia	Wikimedia Foundation	3010 people, 255 military units	API, web pages
16	Knights of the Mannerheim Cross	Knights of the Mannerheim Cross Foundation	191 people, 1120 medal awardings	API, web pages
17	Military historical literature (9 sources)	-	1050 war events, 2900 military units, 585 people	printed text
18	Finnish Spatio-Temporal Ontology	Aalto University	488 polygons of wartime municipalities	RDF
19	AMMO Ontology of Finnish Historical Occupations	Aalto University	3090 occupational labels	RDF

cluding original war diaries (2) are of central importance. Other integrated datasets include, among others, a massive collection of wartime photographs (7), memoirs of soldiers (8), historical maps (10), biographies (12), etc. In addition to people and units, historical (4, 9) and contemporary (11) places, are widely used for data linking. The semantic backbone of WarSampo is the 1050 WW2 events based on military historical literature (17).

4. Data Model

The source datasets of Table 1 were transformed into RDF and harmonized into a coherent whole using an event-based data model. Here the concepts in the source datasets are described using metadata schemas [25], e.g., DCMI Metadata Terms (DCT), and vocabulary models, such as SKOS and RDF Schema (RDFS). This section first motivates the event-based modeling approach used in WarSampo and then presents in more detail the model, core classes, and properties used.[11]

Representing Wars as Events. Since wars are essentially sequences of events, an obvious choice for representing military history is event-based modeling. There are many approaches to modeling events [26–30]. We use CRM with extensions to military historical concepts as the conceptual framework. There are many reasons for this: Firstly, as a strongly event-based model, CRM is suitable for harmonizing the history of wars, Secondly, CRM is an ISO standard (21127:2014), which means that "reinventing the wheel" can be minimized in data modeling. Documen-

[11] The data model is available on GitHub: https://github.com/SemanticComputing/Warsampo-schema.

tation and tooling are readily available for the standard and reuse of the data by others is easier. Thirdly, as CRM describes the real world rather than documents about it, it can be used effectively for harmonizing the heterogeneous source data for a unified representation of the wars and related materials. Using events also makes it possible to describe the changes of status of different entities, such as people and military units. Furthermore, using a common model for all the datasets makes querying the data more uniform.

The used CRM classes and their subclasses are presented in Figure 1 and the used namespace prefixes in Table 2. The class structure was designed and extended iteratively, as the amount of source datasets and links between them increased. In Figure 1, the RDFS subclass relation is represented with a white headed arrow. The relationships between class instances are presented with various properties in the KG, which are divided into two categories based on their certainty: 1) relations that are generated directly from the source dataset information (solid arrows), e.g., a birth event created from a person's birth date in a death record, and 2) relations that are generated using entity linking methods (dotted arrows), e.g., to link a person mentioned in the caption of a photograph. Entity linking methods use heuristics and produce a small amount of erroneous links, which is discussed in Section 6.

Table 2
Namespaces of WarSampo classes and their main properties

Prefix	Namespace
crm	http://www.cidoc-crm.org/cidoc-crm/
rdfs	http://www.w3.org/2000/01/rdf-schema#
skos	http://www.w3.org/2004/02/skos/core#
dct	http://purl.org/dc/terms/
:	http://ldf.fi/schema/warsa/
hipla	http://ldf.fi/schema/hipla/

CRM has an internal way of representing the types of entities, with the property *crm:P2_has_type*. However, the common way of representing specific types in LD is by introducing classes and subclasses for each specific type, and using *rdf:type* to state that a resource is an instance of a class. This approach is used in WarSampo, as it is more expressive, allowing multiple inheritance. In WarSampo, CRM is extended by creating new subclasses for representing the military historical domain. The modeling decision is based on the need to use custom properties for the subclasses, that would not be valid for a whole CRM class. This facilitates interoperability with other systems based on CRM.

Events are represented strictly as subclasses of *crm:E5_Event* depicted on the right in Figure 1. Also the other core classes in the data model are from CRM. However, for some information in the source datasets, modelling them using CRM is not feasible, e.g., marital statuses, or nationalities, as the way to model them with CRM is using groups and events, which is not in line with how people intuitively organize this kind of information [13]. In such cases, the information is annotated using simple SKOS vocabularies.

Literal names of the WarSampo resources are represented using properties *skos:prefLabel* and *skos:altLabel*, instead of the more verbose CRM label appellations, as there is no metadata available about the appellations in the data sources. Information sources are given with the property *dct:source*, and textual descriptions with *dct:description*. The data model can be extended with new CRM subclasses as needed, e.g., when integrating new datasets into the KG.

Core Classes. The WarSampo core classes are presented in Figure 2, with instance and link counts between the class instances. The arrow direction depicts the direction of linking and LOD Cloud refers to the global LOD Cloud. Next, each core class is explained, highlighting its most important properties. Core classes contained within a *domain ontology (DO)* are shown as green rectangles and the RDF *metadatasets (MDS)* using the DOs are shown with yellow rounded rectangles.

Person. (sources 1, 5, 7, 12, 14, 15, 16, 17 in Table 1) The WarSampo person instances have been created [24] from multiple source datasets. The source datasets provide varying levels of detail about people. For most of the people (sources 1 and 14) we have ample biographical metadata, but in some cases the level of detail is not sufficient for disambiguating a person, e.g., only surname and military rank may be known.

The person resources are modeled as instances of *:Person*, a subclass of *crm:E21_Person*. Person resources are further enriched with events created from the source information, e.g., *:Birth*, *:Battle*, *:Death*, *:PersonJoining*, *:Promotion*, or *:MedalAwarding*.

Military Unit. (sources 2, 5, 6, 15, 17) The military unit resources are modeled as instances of *:MilitaryUnit*, a subclass of *crm:E74_Group*. Unit activity is expressed as various related events, e.g., *:Formation*, *:Dissolution*, *:Battle*, and *:TroopMovement*.

During the WW2, changes were made to the army hierarchy: the unit identification codes and unit names were changed occasionally in order to confuse the enemies, and different units have even used identi-

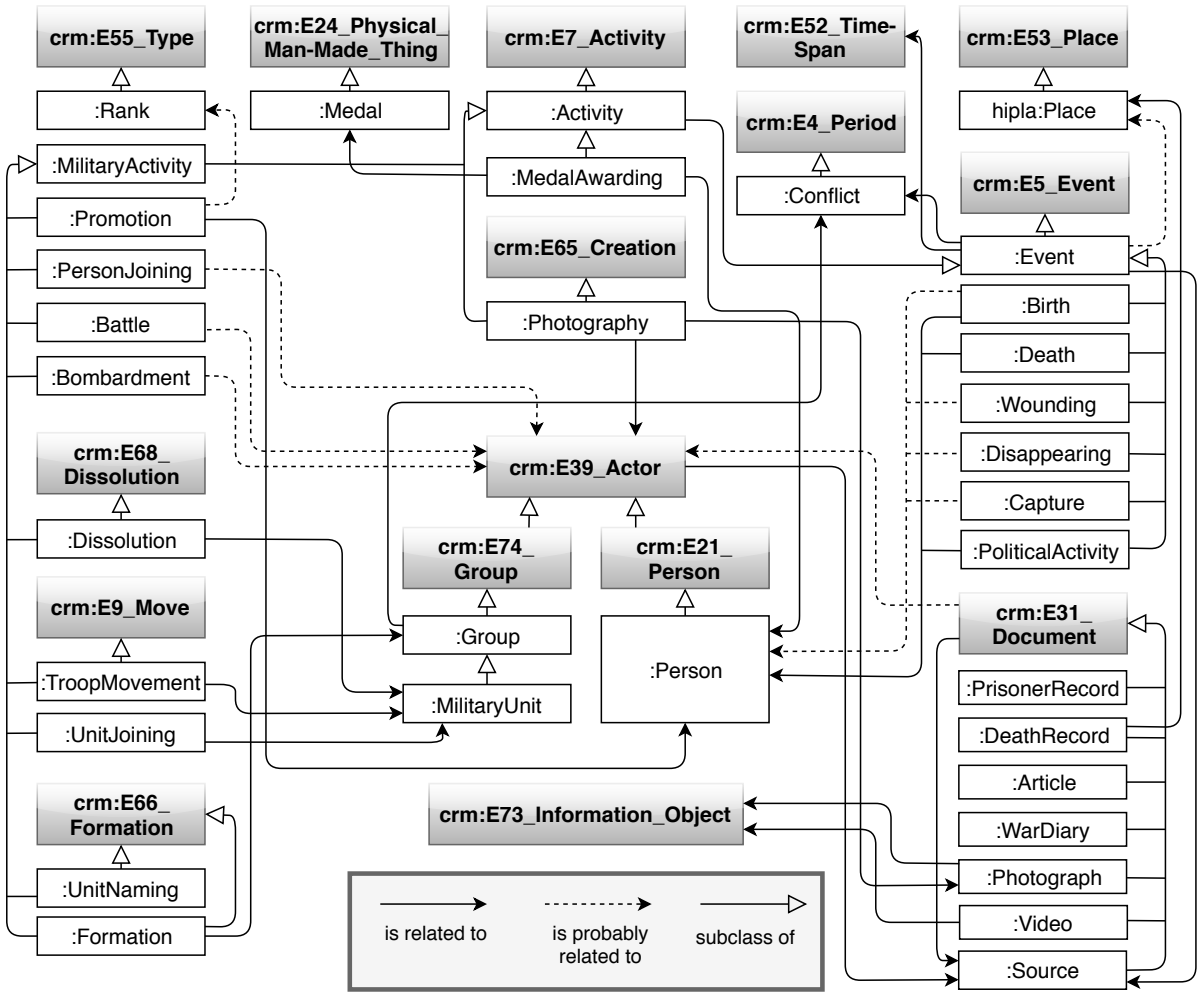

Figure 1. The CRM based WarSampo data model for representing military history as events.

cal names. The army hierarchy, including the temporal changes made in it, is modeled with *:UnitJoining* events that link a unit into its superior unit [24] .

Death Record. (source 1) The death records (DR) contain information about the ca. 94 700 fallen in the Finnish fronts in WW2 [23]. They have served as the primary source of person instances in WarSampo. The data model of person instances is extended based on the DRs, to contain events of wounding and disappearing.

The DRs are modeled as instances of *:DeathRecord*, which is a subclass of *crm:E31_Document*. From each DR, there is a *crm:P70_documents* relation to the corresponding person instance. The DRs are described with custom properties that correspond to the columns of the source spreadsheet, while each DR corresponds to a spreadsheet row. The DR properties convey information about, e.g., the person's occupation, the number of children, marital status, and burial place, using custom SKOS vocabularies. The property values are linked, when possible, to corresponding shared DOs (e.g., Places).

Prisoner Record. (source 14) Prisoner Records (PR) contain information of the ca. 4500 people captured as prisoners of war by the Soviet Union [7]. They are modeled as documents (class *:PrisonerRecord*) similarly as the DRs. Some properties are shared between the PRs and DRs, but in most cases the semantics is different and separate properties are used, that share a common superproperty. Typically, the PR properties depict the person's situation at the time of capture, whereas the DRs depict the situation at the time of death.

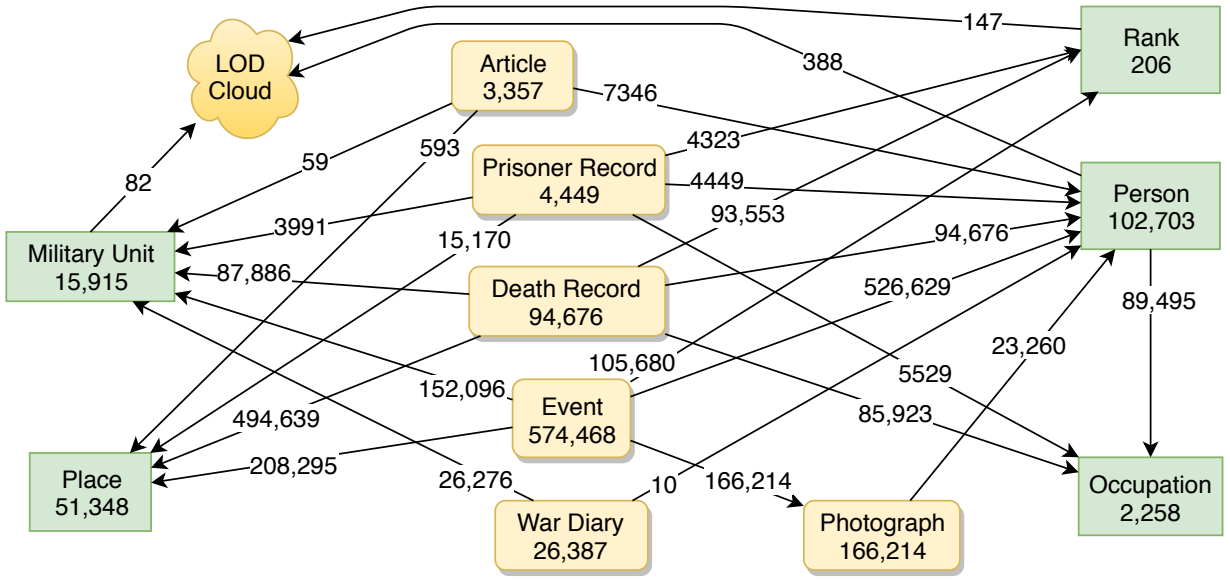

Figure 2. WarSampo core classes with instance counts and linkage between the class instances.

The PRs contribute new person instances and extend the person data model with the capturing events. The PRs often contain multiple values for a property, and source annotations for property values, modeled as RDF reifications.

Event. WarSampo events have been classified into 19 subclasses of the class *crm:E5_Event*, which are shown in Figure 1. They are used to model 1) war events (source 17), e.g., battles and bombardments, 2) political activities (source 17), and 3) events that describe the history of the actors in the war (actor-related sources).

Each event is an instance of *:Event* or one of its subclasses (e.g., *:PoliticalActivity*, *:Battle*, *:Bombardment*). Events are described with textual representations via *dct:description*, time spans, and places of occurrence, if applicable, linking the events to Places DO. The events are linked to actors by several properties, e.g. *crm:P11_had_participant*, *crm:P14_carried_out_by*, and *crm:P100_was_death_of*.

Place. (sources 3, 4, 9, 10, 11, 18) WarSampo employs four distinct types of geographical data: 1) The National Archives' list of counties and municipalities in 1939–1945, enriched with polygon boundaries from the Finnish Spatio-Temporal Ontology[12], 2) Geocoded Karelian map names, 3) War cemeteries, and 4) the current Finnish Place Name Register. In addition, 461 historical map sheets were rectified on modern maps [31].

The geographical data within WarSampo is modeled with a simple schema [32], which contains properties for the place name: coordinates, a polygon, a place type, and part-of relationship of the place. Each place is an instance of a subclass of *crm:E53_Place*. The Finnish Place Name register is used as an external DO, served on a separate endpoint[13].

Photograph. (source 7) WarSampo contains 164 000 wartime photographs with their metadata, taken by Finnish soldiers, as well as 2450 recent photographs of the Finnish war cemeteries. The photographs are represented as instances of the *:Photograph* class. Photography events (class *:Photography*) represent the taking (i.e., creation) of photographs, so that photographs that have been taken the same day and have the same description are grouped in the same event. Modeling the photographs using events has the benefit of making it possible to handle them the same way as other event-based entities and placing them on timelines. Property values link photographs to the DOs of people, military units, and places.

War Diary. (source 2) Metadata of hand-written war diaries are given as instances of the *:WarDiary* class, including *dct:hasFormat* links to the corresponding digitized online documents provided by the Na-

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

[13] http://ldf.fi/pnr/sparql

tional Archives of Finland. The property *crm:P70_documents* links to related military units or people.

Article. (source 8) Metadata of the Kansa Taisteli war veteran magazine articles are given as *:Article* instances. The article metadata is linked to WarSampo DOs of people, military units, and places.

Occupation. (source 19) The AMMO Ontology of Finnish Historical Occupations [22] harmonizes the diverse occupational labels present in the DRs and PRs. AMMO provides the means to study people using social stratification measures via links to the international HISCO [33] classification of occupations, and to another national level classification.

5. Populating the Data Model

The process of creating the WarSampo KG started with the creation of shared DOs [19], shown on the top of Figure 3. The MDSs created from the source datasets, were then linked to the DOs. Some of the early DOs, i.e., 5610 people, military units, military ranks, and medals, have involved manual work, and the processes used to create them are not repeatable. This is also true for person record specific lightweight ontologies used by the death records and the prisoner records. These DOs are maintained directly in RDF and a repeatable data transformation pipeline was built on top of those.

To create a harmonized view of the wars, it is vital to reconcile the entities in the source datasets, by using the shared DOs. In most cases, the reconciliation is based on probabilistic NEL [34], in which a small number of erroneous or missing links is not considered a problem. As a general principle, we have tried to link more rather than less, focusing on recall rather than precision. This enables us to provide at least the relevant links for the users of the data to find more information that they might be interested in. If we emphasized precision more, some relevant information might not be found. We trust in the users' ability to evaluate the links and give feedback if a link is wrong. In some cases, like when disambiguating references to people, we pursued to maximize both recall and precision.

When NEL is used to link literal values to resources, the original values are preserved with a separate property, in order to provide enough information for the user of the data to evaluate whether the generated link might be incorrect.

Transformation Pipeline. A repeatable data transformation pipeline is used for building the majority of the KG from the source datasets. The processes in the pipeline align and transform the source datasets into the WarSampo data model and link entities to the WarSampo DOs.

If the source datasets are updated, the pipeline can be used to update the KG. By recreating the KG, the pipeline also handles change propagation caused by changes in the MDSs or DOs [20, 35], which may cause other parts of the KG to need to adapt to the changes or the linking between resources may become invalid. Several of the data transformation processes employ Docker to increase reproducibility [36].

Figure 3 shows the data transformation pipeline, and links created by the entity linking to the DOs. The boxes represent structured data and the cylinders RDF data, with the yellow color depicting DOs and the green color depicting MDSs. The boxes from which the processes start show the corresponding source numbers from Table 1.

Because of the interlinking between datasets, different change propagation scenarios emerge when updating the source datasets and DOs. The general strategy to best handle the change propagation scenarios is to 1) transform DOs, 2) transform the datasets which both link to the Person DO and create new person instances, and 3) transform datasets that link to the DOs, but do not alter them. The steps shown in Figure 3 are:

1. The place transformation processes convert three source CSVs and one source XML file into RDF, along with the cemetery photograph metadata.
2. The Casualties transformation process transforms the CSV into RDF death records, and links them to the DOs of military ranks, military units, occupations, places, and people [23]. The death records are matched to already existing person instances using probabilistic record linkage [37], with a logistic regression based machine learning implementation. New person instances are created in the Persons DO for the death records that don't match any existing person.
3. The Prisoners of War dataset transformation process [7] is similar to the previous step, and links to the same DOs.
4. The war and political events originate from OCR'd texts, which are then structured into CSV files. This step takes the CSVs as input, transforms them into RDF, and links entities to the DOs [5].
5. Photograph metadata is transformed from CSV into RDF, enriched by images using the data

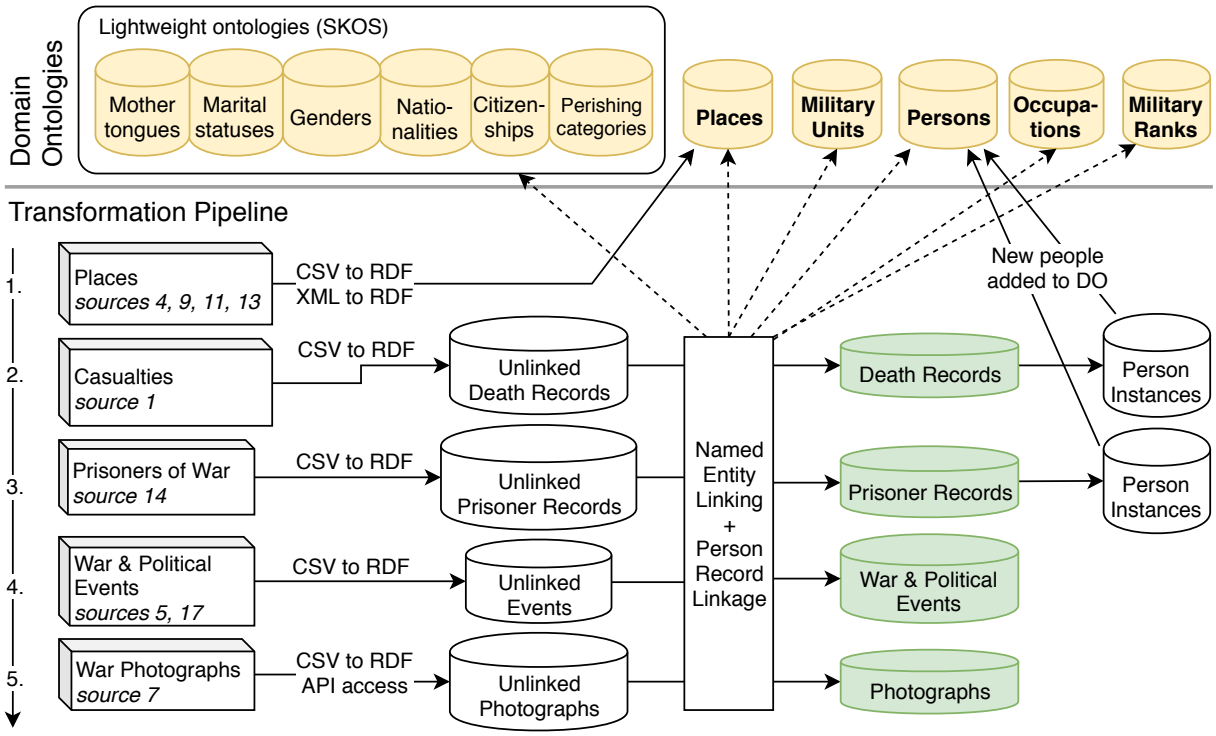

Figure 3. The 5-step WarSampo data transformation process. Dashed arrows represent entity linking, while solid arrows convey data flow.

provider's API, and linked to the DOs of military units, people, and places.

The resulting WarSampo KG consists of 14 300 000 triples, separated into multiple DOs and MDSs.

Data Publication. The KG is available on the Linked Data Finland (LDF) platform [38], providing a home page for the KG [14], and a public SPARQL endpoint[15]. To support reuse, the home page provides additional information about the KG, such as, 1) schema documentation automatically generated by the platform, 2) example SPARQL queries, and 3) metadata as a *SPARQL Service Description*[16], containing *Vocabulary of Interlinked Datasets (VoID)*[17] metadata.

The WarSampo SPARQL endpoint is hosted on an Apache Jena Fuseki[18] SPARQL server. The whole KG and Fuseki are contained in a Docker image, that can be easily built and started when and where needed. The DOs and the transformation pipeline results are separated into individual data repositories, which are included in the image as Git submodules.

The platform provides dereferencing of URIs for both human users and machines, and a generic RDF browser for technical data users, where a user is redirected if a WarSampo URI is visited directly with a web browser. The WarSampo URIs are of the form http://ldf.fi/warsa/DATASET/ID where *DATASET* is the name of the MDS or DO. The *ID* is an identifier consisting of a prefix and a running number, for example: http://ldf.fi/warsa/photographs/sakuva_57717.

The KG is also available in Zenodo, with an associated canonical citation [39]. The KG is licensed with the open Creative Commons BY 4.0 license.

6. Quality of Data

The WarSampo KG is based on the heterogeneous source datasets that are considered being of high quality, since most datasets originate from established national authorities. The data has not been corrected or amended in any way, but only converted into RDF and linked as they are.

[14] The home page of the KG: http://www.ldf.fi/dataset/warsa
[15] The public SPARQL endpoint: http://ldf.fi/warsa/sparql
[16] https://www.w3.org/TR/sparql11-service-description/
[17] https://www.w3.org/TR/void/
[18] https://jena.apache.org/documentation/fuseki2/

The KG adheres to the 5th star level of the 5-star LD publishing principles [40]. In addition, the LDF platform provides an explicit schema and an online documentation[19] to extend the LD publication quality to the sixth star, as suggested in the proposed 7-star model [38]. The data has been validated syntactically by the transformation pipeline and the SPARQL Server. Some schema-based validations regarding selected datasets are underway as the first steps towards obtaining the 7th star; this would require proof that the data conforms to the published schemas. Also some semantic, knowledge-based validation tests were made using SPARQL queries. These tests found out some semantic errors present in the source datasets. For example, there are a few people recorded as being wounded after their death.

Quality of Vocabulary Use. The quality of vocabulary use is on the 4th star level of the five stars of vocabulary use [41]. The WarSampo metadata schema is dereferencable by humans (1 star), and machines (2 stars), it is linked to other vocabularies, e.g., CRM, DCT, and RDFS (3 stars), and it is annotated using DCT, SKOS, and OWL vocabularies (4 stars).

Quality of Entity Linking. The WarSampo entity linking consists of NEL, person record linkage, and a few manually created links.

The NEL of event descriptions to the DOs of people, military units, and places, is accomplished with F_1 scores of 0.88, 1.00, and 0.88, respectively [19]. The NEL of photograph metadata to the DOs of people, military units, and places, is accomplished with F_1 scores of 0.80, 1.00, and 0.77, respectively [19]. The NEL of magazine article metadata to the DOs of military units, and places, is accomplished with F_1 scores of 0.79 and 0.62, respectively [19].

The person record linkage of death records results in 613 death records linked to some of the 5610 pre-existing person instances, while for the remaining 94 100 death records, new person instances are created.

The person record linkage of prisoner records results in 1400 PRs linked to some of the 99 700 pre-existing person instances, while creating 3030 new person instances in the Persons DO.

The precision of the person record linkage of both the death records and prisoner records was manually evaluated to be 1.00, based on randomly selecting 150 links from the total of 620 links for death records, and 200 links from the total of 1400 links for the prisoner records. The information on the person records and the person instances was compared, and all of the records were interpreted to be depicting the same actual people with high confidence.

External Connectivity. Linkage from WarSampo to external resources has been provided to facilitate reuse. WarSampo is connected to the national Finnish ontology infrastructure, by a total of 6110 links, of which 5530 is to KOKO[20], a collection of national core ontologies, and the remaining 582 to YSA[21]. The KOKO linkage contains 3380 keyword annotations of magazine articles and 2140 *skos:relatedMatch* links from AMMO occupation concepts. The YSA links are additional place annotations of the war events that are in geographical scope more global than the WarSampo place ontologies.

There are 3360 external links to the digitized Kansa Taisteli magazine service[22] hosted by the Association for Military History in Finland. There are also 26 400 of external links to the digitized war diaries[23] hosted at the National Archives of Finland.

Linkage to other datasets of the global LOD Cloud[24] consist of 311 links to DBpedia, 159 links to Wikidata, 147 links to Muninn World War I, and 2 links to Cross-Ref DOI Resolver. The military personnel and army units are linked to DBpedia and Wikidata, and the military ranks to Muninn World War I. Additionally, there are 2190 links to Finnish DBpedia.

7. Stability of Data

The KG is mature enough to be relatively static, with only minor error corrections predicted to happen in the near future. New DOs can be added to the ontology infrastructure, without affecting the existing DOs, as the DOs are separated into distinct components, which are handled separately in the transformation pipeline.

The URIs of the Casualties MDS have been revised after initial release, stemming from the MDS originating from a time before the WarSampo infrastructure, and it had URIs which were not in the WarSampo namespace. In 2018, the MDS was revised to

[19] http://ldf.fi/schema/warsa/

[20] KOKO is a collection of Finnish core ontologies, which are merged together: http://finto.fi/koko/en/

[21] YSA is a general thesaurus in Finnish, covering all fields of research and knowledge, containing common terms and geographical names for content description: https://finto.fi/ysa/en/

[22] http://kansataisteli.sshs.fi/

[23] http://digi.narc.fi/digi/dosearch.ka?atun=65.SARK

[24] https://lod-cloud.net/dataset/warsampo

be fully integrated into WarSampo: the namespace was changed, the schema was revised, and the used source dataset was updated. The Casualties transformation process (step 2 in Figure 3) was revised to be fully repeatable and to use person record linkage that is able to adapt to changes in the pre-existing Persons DO. Currently, the used WarSampo URIs can be considered stable.

The KG is versioned using semantic versioning 2.0.0[25], and the KG version discussed in this article is 2.1.0, which includes the prisoners of war dataset, due to be released in November 2019. The full retrospective version history is given in Table 3.

Table 3
WarSampo KG major and minor version history

Version	Published	Description
1.0.0	Nov 2015	Initial public release
1.1.0	Nov 2017	War cemeteries addition
2.0.0	May 2018	Backwards-incompatible URI changes in the Casualties MDS
2.1.0	Nov 2019	Prisoners of war addition

The Linked Data Finland platform, on which the KG is hosted, is actively maintained by the authors of this article and has been operational since 2014.

8. Usefulness

Semantic Portal. The WarSampo Semantic Portal provides end users with nine different WWW based perspectives to the underlying KG. Each perspective is a separate JavaScript application, designed to convey information related to a source dataset or a certain class, in an intuitive and user-friendly way [5]. Instances of core classes, such as people, units, and places, have their "home pages" whose URLs are of the form http://www.sotasampo.fi/en/page?uri=URI, where *URI* is the identifier of the corresponding individual. This makes it easy for the application perspectives or any external application to make reference to WarSampo contents, which facilitates cross-application linking.

The WarSampo KG has been accessed and used by 550 000 end users through the WarSampo Semantic Portal, equivalent to 10% of the population of Finland. We have received written feedback from over 400 end users, mostly through the portal's feedback form. The majority of the feedback contain corrections to the personal information of a respondent's relative. The corrections are stored and supplied to the data providers for further consideration. There is an active open Facebook group[26] for community discussions.

Third-party Use. The core part of KG, the Casualties MDS, has been used as a basis for another popular Finnish WW2 portal, Sotapolku[27], a system aiming at crowdsourcing detailed wartime histories of the Finnish soldiers.

Wikidata has linked some Finnish person instances to WarSampo with a distinct WarSampo property, e.g., the commander-in-chief C. G. E. Mannerheim[28].

Parts of the KG, especially the Places DO and historical maps have been reused in the Finnish historical place and map service Hipla[29] as geo-gazetteers [21] and in the popular NameSampo service[30] for toponomastic research [42].

Finally, the KG was used for enriching data in the external semantic web applications *Norssi High School Alumni* [43], and *BiographySampo* [44].

Known Shortcomings and Future Work. Event-based modeling is an effective approach to representing wars, enabling the harmonization of heterogeneous data, that can be used in spatio-temporal analytics and user interfaces without the need to adjust the queries for each source dataset separately. The downside of using an event-based model for all the datasets is its complexity and verbosity: photographs are, for example, modeled as an image and an event creating it, which can lead to complex and slow queries.

Another problem is data maintenance: data modeled with CRM is considerably difficult to edit directly, due to verbosity and high level of interlinking between resources. Our solution is to support maintenance of the source datasets, which can be repeatedly integrated into the KG using the data transformation pipeline.

The data transformation practices have evolved during the project, and only later datasets are integrated into the KG with repeatable processes. Also modeling conventions have improved, and there are slight variations in how information from different source datasets is modeled.

The transformation pipeline handles most change propagation scenarios, but in some rare cases the initial

[25] https://semver.org/spec/v2.0.0.html

[26] https://www.facebook.com/groups/sotasampo/

[27] http://sotapolku.fi

[28] https://www.wikidata.org/wiki/Q152306

[29] http://hipla.fi

[30] http://nimisampo.fi

DOs need manual updates. For example, if the Places DO changes, the initial state of the Persons DO may need to adapt to the changes, as there are references to e.g., municipalities of birth.

In entity linking, disambiguating some entity types without much context information has been found difficult. For example, place names are usually unambiguous on the municipality level, but automatically disambiguating the names of villages, farms, and lakes can not be done reliably due to high amount of synonymy. Furthermore, place names are often used also as surnames, which is a source of confusion in NEL from free text.

The amount of open, structured, and digitized source datasets about the war is limited. In WarSampo, the focus is on the fallen soldiers, and not much is known about the soldiers who survived the war, except for the high ranking officers who can be considered public figures. In the future, plenty of new material will become available through digitization, raising privacy concerns regarding the people who might still be alive.

9. Conclusion

The WarSampo project has transformed a number of previously isolated source datasets into a harmonized KG. Besides the large number of links between entities, also whole new entities have been extracted from textual content, e.g., people from photograph descriptions, and military units from war diaries.

The WarSampo KG enables queries that were not possible before: for example fetching all WW2 data related to a specific place, or constructing the history of a single soldier based on corresponding military unit information. By publishing shared domain ontologies and data about WW2 for everybody to use in annotations, future interoperability problems can be prevented before they arise.

Acknowledgements

Our work is funded by the Academy of Finland, the Association for Cherishing the Memory of the Dead of the War, the Finnish Ministry of Education and Culture, the Finnish Cultural Foundation, the Memory Foundation for the Fallen, and the Terijoki Trust.

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

References

[1] S. Graham, I. Milligan and S. Weingart, *Exploring big historical data. The historian's macroscope*, Imperial College Press, London, UK, 2015.

[2] A. Burdick, J. Drucker, P. Lunenfeld, T. Presner and J. Schnapp, *Digital Humanities*, The MIT Press, 2012.

[3] R.M. Citino, Military Histories Old and New: A Reintroduction, *The American Historical Review* **112**(4) (2007), 1070–1090. http://www.jstor.org/stable/40008444.

[4] C. Bizer, T. Heath and T. Berners-Lee, Linked Data–The Story So Far, *Semantic services, interoperability and web applications: emerging concepts* (2009), 205–227.

[5] E. Hyvönen, E. Heino, P. Leskinen, E. Ikkala, M. Koho, M. Tamper, J. Tuominen and E. Mäkelä, 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)*, Springer, 2016, pp. 758–773.

[6] E. Ikkala, M. Koho, E. Heino, P. Leskinen, E. Hyvönen and T. Ahoranta, Prosopographical Views to Finnish WW2 Casualties Through Cemeteries and Linked Open Data, in: *Proceedings of the Workshop on Humanities in the Semantic Web (WHiSe II)*, CEUR Workshop Proceedings, 2017.

[7] M. Koho, E. Ikkala and E. Hyvönen, Reassembling the Lives of Finnish Prisoners of the Second World War on the Semantic Web, CEUR Workshop Proceedings, 2019, In press.

[8] R. Hoekstra, A. Meroño-Peñuela, K. Dentler, A. Rijpma, R. Zijdeman and I. Zandhuis, An Ecosystem for Linked Humanities Data, in: *The Semantic Web*, Springer International Publishing, Cham, 2016, pp. 425–440.

[9] A. Meroño-Peñuela, A. Ashkpour, M. Van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach and F. Van Harmelen, Semantic technologies for historical research: A survey, *Semantic Web* **6**(6) (2015), 539–564.

[10] V. de Boer, A. Meroño-Peñuela and C.J. Ockeloen, Linked Data for Digital History: Lessons Learned from Three Case Studies, *Anejos de la Revista de Historiografía* (2016), 139–162.

[11] G. Nagypál, R. Deswarte and J. Oosthoek, Applying the semantic web: The VICODI experience in creating visual contextualization for history, *Literary and Linguistic Computing* **20**(3) (2005), 327–349.

[12] M. Doerr, The CIDOC CRM – an Ontological Approach to Semantic Interoperability of Metadata, *AI Magazine* **24**(3) (2003), 75–92.

[13] E. Mäkelä, J. Törnroos, T. Lindquist and E. Hyvönen, WW1LOD - An application of CIDOC-CRM to World War 1 Linked Data, *International Journal on Digital Libraries* (2016).

[14] T. Collins, P. Mulholland and Z. Zdrahal, Semantic Browsing of Digital Collections, in: *Proceedings of the 4th International Semantic Web Conference (ISWC 2005)*, Springer, 2005, pp. 127–141.

[15] V. de Boer, J. van Doornik, L. Buitinck, M. Marx and T. Veken, Linking the kingdom: enriched access to a historiographical text, in: *Proceedings of the 7th International Conference on Knowledge Capture (KCAP 2013)*, ACM, 2013, pp. 17–24.

[16] A. van Nispen and L. Jongma, Holocaust and World War Two Linked Open Data Developments in the Netherlands, *Umanistica Digitale* **3**(4) (2019). doi:10.6092/issn.2532-8816/9048.

[17] R. Sprugnoli, G. Moretti and S. Tonelli, LOD Navigator: Tracing Movements of Italian Shoah Victims, *Umanistica Digitale* **3**(4) (2019). doi:10.6092/issn.2532-8816/9050.

[18] E. Hyvönen, J. Tuominen, E. Mäkelä, J. Dutruit, K. Apajalahti, E. Heino, P. Leskinen and E. Ikkala, Second World War on the Semantic Web: The WarSampo Project and Semantic Portal, in: *Proceedings of the ISWC 2015 Posters & Demonstrations Track*, CEUR Workshop Proceedings, 2015, Vol 1486.

[19] E. Heino, M. Tamper, E. Mäkelä, P. Leskinen, E. Ikkala, J. Tuominen, M. Koho and E. Hyvönen, Named Entity Linking in a Complex Domain: Case Second World War History, in: *Language, Technology and Knowledge*, Springer, 2017.

[20] M. Koho, E. Ikkala, E. Heino and E. Hyvönen, Maintaining a Linked Data Cloud and Data Service for Second World War History, in: *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. 7th International Conference, EuroMed 2018, Nicosia, Cyprus*, Vol. 11196, Springer-Verlag, 2018.

[21] E. Ikkala, E. Hyvönen and J. Tuominen, An Ontology of World War II Places for Linking and Enriching Heterogeneous Historical Data Sources, in: *Abstracts, 17th International Conference of Historical Geographers (ICHG 2018), No. 194*, 2018.

[22] M. Koho, L. Gasbarra, J. Tuominen, H. Rantala, I. Jokipii and E. Hyvönen, AMMO Ontology of Finnish Historical Occupations, in: *Proceedings of the the 1st International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH'19)*, CEUR Workshop Proceedings, 2019, pp. 91–96, Vol 2375.

[23] M. Koho, E. Hyvönen, E. Heino, J. Tuominen, P. Leskinen and E. Mäkelä, Linked Death - Representing, Publishing, and Using Second World War Death Records as Linked Open Data, in: *The Semantic Web: ESWC 2017 Satellite Events*, Springer-Verlag, 2017, pp. 369–383.

[24] P. Leskinen, M. Koho, E. Heino, M. Tamper, E. Ikkala, J. Tuominen, E. Mäkelä and E. Hyvönen, Modeling and Using an Actor Ontology of Second World War Military Units and Personnel, in: *Proceedings of the 16th International Semantic Web Conference (ISWC 2017)*, Springer-Verlag, 2017, pp. 280–296. https://doi.org/10.1007/978-3-319-68204-4_27.

[25] M.L. Zeng and J. Qin, *Metadata*, 2nd edn, ALA Neal-Schuman, USA, 2016.

[26] M. Rovera, 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*, Springer, 2016, pp. 845–852.

[27] Y. Raimond, S.A. Abdallah, M.B. Sandler and F. Giasson, The Music Ontology, in: *ISMIR 2007: Proceedings of the 8th International Conference on Music Information Retrieval*, Austrian Computer Society, 2007.

[28] A. Scherp, T. Franz, C. Saathoff and S. Staab, F–a Model of Events Based on the Foundational Ontology Dolce+DnS Ultralight, in: *Proceedings of the Fifth International Conference on Knowledge Capture*, K-CAP '09, ACM, New York, NY, USA, 2009, pp. 137–144. ISBN 978-1-60558-658-8.

[29] R. Shaw, R. Troncy and L. Hardman, LODE: Linking Open Descriptions of Events, in: *The Semantic Web*, Springer Berlin Heidelberg, 2009, pp. 153–167.

[30] W.R. van Hage, V. Malaisé, R. Segers, L. Hollink and G. Schreiber, Design and use of the Simple Event Model (SEM), *Journal of Web Semantics* **9**(2) (2011), 128–136.

[31] E. Ikkala, E. Hyvönen and J. Tuominen, Geocoding, Publishing, and Using Historical Places and Old Maps in Linked Data Applications, in: *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference*, CEUR Workshop Proceedings, 2018, pp. 228–234.

[32] E. Hyvönen, E. Ikkala and J. Tuominen, Linked Data Brokering Service for Historical Places and Maps, in: *Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe)*, CEUR Workshop Proceedings, 2016, pp. 39–52, Vol 1608.

[33] M.H.D. Van Leeuwen, I. Maas and A. Miles, *HISCO: Historical International Standard Classification of Occupations*, Leuven University Press, 2002.

[34] B. Hachey, W. Radford, J. Nothman, M. Honnibal and J.R. Curran, Evaluating Entity Linking with Wikipedia, *Artificial Intelligence* **194** (2013), 130–150.

[35] L. Stojanovic, A. Maedche, B. Motik and N. Stojanovic, User-driven ontology evolution management, in: *International Conference on Knowledge Engineering and Knowledge Management*, Springer, 2002, pp. 285–300.

[36] J. Cito, V. Ferme and H.C. Gall, Using Docker Containers to Improve Reproducibility in Software and Web Engineering Research, in: *Web Engineering*, Springer International Publishing, 2016, pp. 609–612.

[37] L. Gu, R. Baxter, D. Vickers and C. Rainsford, Record Linkage: Current Practice and Future Directions, *CSIRO Mathematical and Information Sciences* (2003), CMIS Technical Report No. 03/83.

[38] E. Hyvönen, J. Tuominen, M. Alonen and E. Mäkelä, 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*, Springer, 2014, pp. 226–230.

[39] M. Koho, E. Heino, P. Leskinen, E. Ikkala, M. Tamper, K. Apajalahti, J. Tuominen, E. Mäkelä and E. Hyvönen, WarSampo Knowledge Graph [Data set], Zenodo, 2019. https://doi.org/10.5281/zenodo.3431121.

[40] T. Berners-Lee, Linked Data - Design Issues, 2006, Accessed: 2019-09-10. http://w3.org/DesignIssues/LinkedData.html.

[41] K. Janowicz, P. Hitzler, B. Adams, D. Kolas, I. Vardeman et al., Five Stars of Linked Data Vocabulary Use, *Semantic Web* **5**(3) (2014), 173–176.

[42] E. Ikkala, J. Tuominen, J. Raunamaa, T. Aalto, T. Ainiala, H. Uusitalo and E. Hyvönen, NameSampo: A Linked Open Data Infrastructure and Workbench for Toponomastic Research, in: *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Geospatial Humanities*, GeoHumanities'18, ACM, 2018, pp. 2:1–2:9. ISBN 978-1-4503-6032-6.

[43] E. Hyvönen, P. Leskinen, E. Heino, J. Tuominen and L. Sirola, Reassembling and Enriching the Life Stories in Printed Biographical Registers: Norssi High School Alumni on the Semantic Web, in: *Proceedings, Language, Technology and Knowledge (LDK 2017)*, Springer-Verlag, 2017, pp. 113–119.

[44] E. Hyvönen, P. Leskinen, M. Tamper, H. Rantala, E. Ikkala, J. Tuominen and K. Keravuori, BiographySampo – Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research, in: *Proceedings of the 16th Extended Semantic Web Conference (ESWC 2019)*, Springer-Verlag, 2019.