Computer Science

Representing, Using, and Maintaining Military Historical Linked Data on the Semantic Web

Mikko Koho

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 71/2020

Representing, Using, and Maintaining Military Historical Linked Data on the Semantic Web

Mikko Koho

The public defense on 15th May 2020 at 12:00 will be organized via remote technology.

Link: https://aalto.zoom.us/j/65732236860

Zoom Quick Guide: https://www.aalto.fi/en/services/zoom-quick-guide

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, https://aalto.zoom.us/j/65732236860, on 15 May 2020 at 12 noon.

Aalto University
School of Science
Department of Computer Science
Semantic Computing Research Group

Supervising professor
Professor Eero Hyvönen, Aalto University & University of Helsinki, Finland

Thesis advisors
Professor Eetu Mäkelä, University of Helsinki & Aalto University, Finland
Doctor Jouni Tuominen, Aalto University & University of Helsinki, Finland

Preliminary examiners
Professor Jose Emilio Labra Gayo, University of Oviedo, Spain
Professor Marcia Lei Zeng, Kent State University, USA

Opponent
Professor Jose Emilio Labra Gayo, University of Oviedo, Spain

Aalto University publication series
DOCTORAL DISSERTATIONS 71/2020

ISBN 978-952-60-3868-1 (printed)
ISBN 978-952-60-3869-8 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-3869-8

Images: Cover Image by Risto Eliel William Orko, source SA-kuva,
Creative Commons BY 4.0

Unigrafia Oy
Helsinki 2020

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Mikko Koho

Name of the doctoral dissertation
Representing, Using, and Maintaining Military Historical Linked Data on the Semantic Web

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 71/2020

Field of research Semantic Web

Manuscript submitted 12 December 2019 **Date of the defence** 15 May 2020

Permission for public defence granted (date) 17 April 2020 **Language** English

□ **Monograph** ☒ **Article dissertation** □ **Essay dissertation**

Abstract

The Second World War is the largest global tragedy in human history. It is extensively documented in historical sources, but this information is scattered in various organizations and countries, written in multiple languages, and represented in heterogeneous formats. Semantic Web technologies provide solutions for combining heterogeneous distributed historical information. By combining information from distributed sources it is possible to get a deeper understanding about the history than by studying the sources individually.

This thesis explores the use of Semantic Web technologies for representing and modeling heterogeneous military historical information as Linked Data, with a focus on depicting the history of Finland in the Second World War. Harmonization and integration of military historical data from distributed sources are studied, while also investigating how to search, browse, analyze, and visualize the resulting Linked Data on web-based user interfaces. Maintenance of the highly interlinked set of graphs exposes new challenges and a solution to tackle them is presented. These topics are studied in the context of building the WarSampo information system.

Linked Data and the event-based CIDOC Conceptual Reference Model are used together in WarSampo to achieve the interoperability of heterogeneous military historical datasets. Events are used as the glue which combines together information from various source datasets. The event-based modeling enables depicting the national military history narrative as data, which can be further enriched with the events of individual military units and soldiers. This idea is demonstrated in the WarSampo semantic portal, which consists of nine different perspectives on the data integrated from distributed sources. Each perspective provides a customized user interface for a certain part of the WarSampo knowledge graph, like war events, persons, wartime photographs, and places.

The knowledge graph is published as open data and is a part of the global Linked Open Data Cloud. The WarSampo portal at http://sotasampo.fi is a popular service for citizens to study the wars, and to find out what happened to their relatives. It has been used by more than 660 000 end users, equivalent to more than 10% of the population of Finland. The proposed methods and data models are useful beyond the geographical and temporal scopes of this research. The aspiration behind the WarSampo project is that by making military historical data more accessible, our understanding about the reality of the war will improve, which also promotes peace in the future.

Keywords semantic web, linked data, interoperability, semantic reconciliation, semantic disambiguation, information systems, military history, digital humanities

ISBN (printed) 978-952-60-3868-1 **ISBN (pdf)** 978-952-60-3869-8

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2020

Pages 214 **urn** http://urn.fi/URN:ISBN:978-952-60-3869-8

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Mikko Koho

Väitöskirjan nimi
Representing, Using, and Maintaining Military Historical Linked Data on the Semantic Web

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietotekniikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 71/2020

Tutkimusala Semanttinen web

Käsikirjoituksen pvm 12.12.2019 **Väitöspäivä** 15.05.2020

Väittelyluvan myöntämispäivä 17.04.2020 **Kieli** Englanti

□ **Monografia** ☒ **Artikkeliväitöskirja** □ **Esseeväitöskirja**

Tiivistelmä

Toinen maailmansota on ihmiskunnan historian suurin globaali tragedia. Se on kattavasti dokumentoitu historiallisissa lähteissä, mutta tämä tieto on hajallaan monissa eri maissa ja niiden sisällä useissa erillisissä organisaatioissa, kirjoitettuna useilla kielillä ja vaihtelevissa formaateissa. Semanttisen webin teknologiat mahdollistavat epäyhtenäisen historiallisen tiedon yhdistämisen useista erillisistä lähteistä. Tiedon yhdistäminen useista lähteistä auttaa ymmärtämään historiaa syvällisemmin kuin tarkastelemalla yksittäisiä lähteitä.

Tämä väitöskirja tutkii semanttisen webin teknologioiden käyttöä sotahistorian esittämiseen ja mallintamiseen linkitettynä datana. Keskiössä on Suomen historian esittäminen toisen maailmansodan ajalta sekä sotahistoriallisen tiedon harmonisointi ja yhdistäminen erillisistä lähteistä. Tutkimuksessa selvitetään, miten syntyvää linkitettyä dataa voidaan hakea, selata, analysoida ja visualisoida web-pohjaisissa käyttöliittymissä ja miten voimakkaasti yhteenlinkittyneitä linkitetyn datan graafeja voidaan ylläpitää. Tätä tutkimusta on tehty osana Sotasampo-tietojärjestelmän kehitystä.

Sotasammossa hyödynnetään Linkitettyä Dataa ja tapahtumapohjaista CIDOC Conceptual Reference Model -tietomallia epäyhtenäisten sotahistoriallisten aineistojen yhteentoimivuuden saavuttamiseksi. Tapahtumat toimivat liimana, joka yhdistää tietoa useista lähdeaineistoista. Tapahtumapohjainen mallintaminen mahdollistaa kansallisen sotahistoriallisen narratiivin esittämisen datana, jota voidaan rikastaa yksittäisiin joukko-osastoihin ja henkilöihin liittyvillä tapahtumilla. Tätä ideaa demonstroidaan Sotasampo-portaalissa, joka sisältää yhdeksän erilaista perspektiiviä eri lähteistä yhdistettyyn tietämysgraafiin. Jokainen perspektiivi tarjoaa käyttöliittymän, joka on räätälöity tiettyyn osaan Sotasammon tietämysgraafista, kuten sodanajan tapahtumiin, henkilöihin, sodanajan valokuviin, tai paikkoihin.

Sotasammon tietämysgraafi on julkaistu avoimena datana ja se muodostaa osan globaalista linkitetyn avoimen datan LOD Cloud -datapilvestä. Avoin Sotasampo-portaali http://sotasampo.fi on suosittu palvelu kansalaisille sota-aineistojen tutkimiseen ja sukulaistensa sotataipaleen selvittämiseen. Sillä on ollut yli 660 000 käyttäjää, joka vastaa yli kymmenesosaa suomalaisista. Esitetyt menetelmät ja tietomallit ovat käyttökelpoisia myös tätä tutkimusta laajemmalla maantieteellisellä ja ajallisella rajauksella. Sotasampo-projektin taustalla on ajatus siitä, että sotahistoriallisen tiedon tuominen helpommin saataville lisää ymmärrystä sodasta ja osaltaan edistää rauhaa tulevaisuudessa.

Avainsanat semanttinen web, linkitetty data, yhteentoimivuus, semanttinen yhteensovittaminen, semanttinen yksikäsitteistäminen, tietojärjestelmät, sotahistoria, digitaaliset ihmistieteet

ISBN (painettu) 978-952-60-3868-1 **ISBN (pdf)** 978-952-60-3869-8

ISSN (painettu) 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2020

Sivumäärä 214 **urn** http://urn.fi/URN:ISBN:978-952-60-3869-8

Preface

The research presented in this thesis was conducted at the Semantic Computing Research Group (SeCo) at the Department of Computer Science, Aalto University, in collaboration with the HELDIG – Helsinki Centre for Digital Humanities at the University of Helsinki.

Many thanks to my thesis supervisor, Professor Eero Hyvönen, for providing this opportunity and for great support. I would like to express my deepest appreciation to my advisors Professor Eetu Mäkelä and Doctor Jouni Tuominen for providing invaluable guidance and support throughout this journey. I also wish to thank the pre-examiners, Professor Jose Emilio Labra Gayo and Professor Marcia Lei Zeng, for valuable feedback.

I'm extremely grateful to Esko Ikkala and Erkki Heino for productive and enjoyable co-operation in the WarSampo project and other projects. I would like to extend my sincere thanks to the other co-authors of the publications of this thesis, Petri Leskinen, Minna Tamper, Lia Gasbarra, Heikki Rantala, Tomi Ahoranta, and Ilkka Jokipii. I'd like to acknowledge the effort of Jérémie Dutruit in doing some groundwork in the WarSampo project that my research could build on.

During this journey I also had great pleasure of working with Babatunde Anafi, Kasper Apajalahti, Matias Frosterus, Pejam Hassanzadeh, Aleksandra Konovalova, Alex Kourijoki, Rafael Leal Tomás, Goki Miyakita, Arttu Oksanen, Sisko Pajari, and Sami Sarsa. Special thanks to Katri Miettinen, Tiia Moilanen, Reijo Nikkilä, and Pertti Suominen, for collaboration regarding the prisoners of war register. Thanks should also go to Timo Hakala, Hanna Hyvönen, Ohto Manninen, Jyrki Tiittanen, and Susanna Ånäs, for their valuable contributions to the WarSampo project.

I wish to thank the organizations supporting the WarSampo project by sharing their data: the National Archives of Finland, the Finnish Defence Forces, the Association for Military History in Finland, Bonnier Publishing, the National Land Survey of Finland, the Finnish Literature Society, the Association of Finnish Camera Clubs, the National Prisoners of War Project, Wikimedia Foundation, and Knights of the Mannerheim Cross Foundation. Lastly, I thank CSC – IT Center for Science, Finland,

for providing computational resources for the research.

The work presented in this thesis has been funded by the Ministry of Education and Culture, the Memory Foundation for the Fallen, the Academy of Finland, the Association for Cherishing the Memory of the Dead of the War, the Association of Finnish Camera Club, and the Finnish Cultural Foundation.

I have received a grant for the doctoral research from Teri-Säätiö (2019) and two travel grants (2016) from the Helsinki Doctoral Education Network in Information and Communications Technology (HICT).

I am extremely grateful to my parents for their support and encouragement. Finally, I sincerely thank my wife Promila for love and support.

Helsinki, April 23, 2020,

Mikko Koho

Contents

Preface 1

Contents 3

List of Publications 5

Author's Contribution 9

Abbreviations 13

1. Introduction 15

1.1 Background and Research Environment 15
1.2 Objectives and Scope 17
1.3 Research Process and Dissertation Structure 19

2. Theoretical Foundation 21

2.1 Military History 21
2.2 Semantic Reconciliation 25
2.3 Semantic Disambiguation and Entity Linking 30
2.4 Web Portals 32
2.5 Maintaining Linked Data 36

3. Results 39

3.1 Modeling and Representing Military History 39
3.2 Harmonizing Heterogeneous Data 44
3.3 Semantic Portal For Military History 47
3.4 Maintaining Military Historical Linked Data 53
3.5 Results Summary 55

4. Discussion 59

4.1 Theoretical Implications 59
4.2 Practical Implications 60
4.3 Reliability and Validity 61

4.4 Recommendations for Further Research 63

Bibliography **65**

Publications **83**

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Mikko Koho, Esko Ikkala, Petri Leskinen, Minna Tamper, Jouni Tuominen, and Eero Hyvönen. WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data. Submitted to *Semantic Web – Interoperability, Usability, Applicability: Special Issue on Semantic Web for Cultural Heritage*, October 2019.

II Petri Leskinen, Mikko Koho, Erkki Heino, Minna Tamper, Esko Ikkala, Jouni Tuominen, Eetu Mäkelä, and Eero Hyvönen. Modeling and Using an Actor Ontology of Second World War Military Units and Personnel. In *The Semantic Web – ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part II*, Claudia d'Amato, Miriam Fernandez, Valentina Tamma, Freddy Lecue, Philippe Cudré-Mauroux, Juan Sequeda, Christoph Lange, and Jeff Heflin (editors), Information Systems and Applications, incl. Internet/Web, and HCI, volume 10588, pages 280–296, ISBN 9783319682037, Springer, Cham, October 2017.

III Mikko Koho, Lia Gasbarra, Jouni Tuominen, Heikki Rantala, Ilkka Jokipii, and Eero Hyvönen. AMMO Ontology of Finnish Historical Occupations. In *Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH 2019), Rome, Italy, June 3, 2019*, Antonella Poggi (editor), CEUR Workshop Proceedings, volume 2375, pages 91–96, ISSN 16130073, online CEUR-WS.org/Vol-2375/short2.pdf, June 2019.

IV Eero Hyvönen, Erkki Heino, Petri Leskinen, Esko Ikkala, Mikko Koho, Minna Tamper, Jouni Tuominen, and Eetu 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: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29 – June 2, 2016, Proceedings*, Harald Sack, Eva Blomqvist, Mathieu d'Aquin, Chiara Ghidini, Simone Paolo Ponzetto, and Christoph Lange (editors), Lecture Notes in Computer Science, volume 9678, pages 758–773, ISBN 9783319341286, Springer, Cham, May–June 2016.

V Esko Ikkala, Mikko Koho, Erkki Heino, Petri Leskinen, Eero Hyvönen, and Tomi 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), Vienna, Austria, October 22, 2017*, Alessandro Adamou, Enrico Daga, and Leif Isaksen (editors), CEUR Workshop Proceedings, volume 2014, pages 45–56, ISSN 16130073, online CEUR-WS.org/Vol-2014/paper-06.pdf, October 2017.

VI Mikko Koho, Esko Ikkala, and Eero Hyvönen. Reassembling the Lives of Finnish Prisoners of the Second World War on the Semantic Web. Accepted for publication in *Proceedings of the Third Conference on Biographical Data in the Digital Age (BD 2019), Varna, Bulgaria*, CEUR Workshop Proceedings, 9 pages, in press, September 2019.

VII Mikko Koho, Erkki Heino, and Eero Hyvönen. SPARQL Faceter—Client-side Faceted Search Based on SPARQL. In *Joint Proceedings of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop co-located with the 13th Extended Semantic Web Conference ESWC 2016, Heraklion, Crete, Greece, May 30, 2016*, Raphaël Troncy, Ruben Verborgh, Lyndon Nixon, Thomas Kurz, Kai Schlegel, and Miel Vander Sande (editors), CEUR Workshop Proceedings, volume 1615, ISSN 16130073, online CEUR-WS.org/Vol-1615/semdevPaper5.pdf, May 2016.

VIII Mikko Koho, Eero Hyvönen, Erkki Heino, Jouni Tuominen, Petri Leskinen, and Eetu Mäkelä. Linked Death — Representing, Publishing, and Using Second World War Death Records as Linked Open Data. In *The Semantic Web: ESWC 2017 Satellite Events: ESWC 2017 Satellite Events, Portorož, Slovenia, May 28 – June 1, 2017, Revised Selected Papers*, Eva Blomqvist, Katja Hose, Heiko Paulheim, Agnieszka Ławrynowicz, Fabio Ciravegna, and Olaf Hartig (editors), Lecture Notes in Computer Science, volume 10577, pages 369–383, ISBN 9783319704067, Springer, Cham, May–June 2017.

IX Mikko Koho, Esko Ikkala, Erkki Heino, and Eero 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, October 29–November 3, 2018, Proceedings, Part I*, Marinos Ioannides, Eleanor Fink, Rafaella Brumana, Petros Patias, Anastasios Doulamis, João Martins, and Manolis Wallace (editors), Information Systems and Applications, incl. Internet/Web, and HCI, volume 11196, pages 138–149, ISBN 9783030017613, Springer, Cham, October–November 2018.

Author's Contribution

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

The author is the lead author of the publication and wrote most of it. The author was the primary developer in designing and implementing the programmatic integration processes of the prisoners of war register and the casualties register into the WarSampo infrastructure, and designed the related data model extensions. The author analyzed the linking quality of these datasets. The author was one of the three primary designers of the WarSampo data transformation pipeline.

Publication II: "Modeling and Using an Actor Ontology of Second World War Military Units and Personnel"

The author contributed to the writing of the publication as a co-author. The author was the primary developer designing and implementing the programmatic integration of the casualties register into WarSampo, providing most of the actors in the actor ontology. The author contributed to the actor ontology schema.

Publication III: "AMMO Ontology of Finnish Historical Occupations"

The author is the lead author of the publication and wrote most of it. The author designed the ontology model, the ontology engineering process, and was in charge of the technical implementation of the ontology.

Publication IV: “WarSampo Data Service and Semantic Portal for Publishing Linked Open Data about the Second World War History”

The author contributed to the writing of the publication as a co-author. The author was the primary developer in designing and implementing the programmatic integration of the casualties register into WarSampo. The author contributed significantly to the design and implementation of the casualties perspective of the WarSampo portal.

Publication V: “Prosopographical Views to Finnish WW2 Casualties Through Cemeteries and Linked Open Data”

The author contributed significantly to the writing of the publication as a co-author. The author implemented the linking of the death records to the war cemetery data. The author was the primary developer in designing and implementing of the visualizations in the WarSampo casualties perspective.

Publication VI: “Reassembling the Lives of Finnish Prisoners of the Second World War on the Semantic Web”

The author is the lead author of the publication and wrote most of it. The author was the primary developer in designing and implementing the integration process of the prisoners of war register into the WarSampo infrastructure. The author was the primary developer in implementing the prisoners of war perspective of the WarSampo portal. The author contributed considerably to the design and implementation of the reshaping of the WarSampo persons perspective.

Publication VII: “SPARQL Faceter—Client-side Faceted Search Based on SPARQL”

The author is the lead author of the publication and wrote most of it. The author formulated the design requirements of the SPARQL Faceter tool and evaluated the tool against the requirements. The author contributed to the technical design and implementation of the tool. The author contributed significantly to the design and implementation of the WarSampo casualties perspective.

Publication VIII: "Linked Death — Representing, Publishing, and Using Second World War Death Records as Linked Open Data"

The author is the lead author of the publication and wrote most of it. The author was the primary developer in designing and implementing the programmatic integration of the casualties register into WarSampo. The author contributed significantly to the design and implementation of the casualties perspective of the WarSampo portal. The author implemented the study of different use cases of the data.

Publication IX: "Maintaining a Linked Data Cloud and Data Service for Second World War History"

The author is the lead author of the publication and wrote most of it. The author contributed considerably to the analysis of change propagation scenarios in WarSampo. The author was the primary developer in designing and implementing the integration process of the prisoners of war register into the WarSampo infrastructure.

In addition to the aforementioned publications, the thesis contains references to related work by the author concerning military history and WarSampo. The first WarSampo publication depicted an early state of the WarSampo dataset and semantic portal [91]. A case study has been made on collaborating with domain experts to integrate a dataset about the Finnish prisoners of war into WarSampo [108]. Named entity linking within WarSampo context has been studied in [83]. A study compared an early version of the faceted search interface of WarSampo's casualties perspective with two other faceted search implementations [125]. A new addition to the WarSampo ontology infrastructure is the work done in harmonizing early 20th century Finnish occupational labels into an ontology with linked social stratification information [67]. The WarSampo knowledge graph discussed in this thesis is published as a dataset with a canonical citation [109]. The Linked Open Data portal of Finnish War Victims in 1914–1922 is presented in [172, 171].

During the research leading up to the thesis, the author has also been involved in other research related to digital humanities and applying computational methods in the cultural heritage domain. These include a project concerned with building a Linked Open Data database of Finnish archaeological finds [219, 220, 198], and another project integrating and harmonizing metadata of pre-modern manuscripts into a Linked Open Data service and portal for manuscript studies [37, 92]. In addition, the author has been part of a project studying text classification and distant reading of historical newspapers [58], and creating a search service for news content employing semantics, topic modeling, and relevance feedback [110].

Abbreviations

API Application Programming Interface

CDEC Jewish Contemporary Documentation Center

CENDARI Collaborative European Digital Archival Research Infrastructure

CIDOC International Committee for Documentation

CIDOC CRM CIDOC Conceptual Reference Model

CRM *see CIDOC CRM*

DC Dublin Core

DCT DCMI Metadata Terms

DO Domain ontology

EDM Europeana Data Model

EHRI European Holocaust Research Infrastructure

FOAF Friend of a Friend

HISCO Historical International Standard of Classification of Occupations

HTML Hypertext Markup Language

HTTP Hypertext Transfer Protocol

IRI Internationalized Resource Identifier

ISO International Organization for Standardization

LDC Linked Data Cloud

LOD Linked Open Data

LODLAM Linked Open Data in Libraries, Archives, and Museums

MDS Metadataset

NEL Named Entity Linking

NER Named Entity Recognition

PDF Portable Document Format

RDF Resource Description Framework

RDFS RDF Schema

SHACL Shapes Constraint Language

ShEx Shape Expressions

SKOS Simple Knowledge Organization System

SPARQL SPARQL Protocol and RDF Query Language

URI Uniform Resource Identifier

VICODI Visual Contextualization of Digital Content

WW1 First World War

WW2 Second World War

XML Extensible Markup Language

1. Introduction

1.1 Background and Research Environment

The Second World War (WW2) has been studied extensively in military historical research [142], and for its vast dimensions, it is considered a clear demonstration of the capacity of human beings for destroying each other and themselves [218]. As more data is becoming available, the possibilities of research applying computational methods to military historical data are increasing. The WW2 is of great interest not only to historians, but to potentially hundreds of millions of citizens globally, whose relatives participated in the war, creating a global shared trauma. However, data about the WW2 is hard to get since it is scattered in various organizations and countries, written in multiple languages, and represented in heterogeneous formats. Combining information from the distributed historical sources supports getting a deeper understanding about the history than by studying the sources individually.

Plenty of information about WW2 exists around the world in Cultural Heritage memory institutions. Most of this information exists only in paper format although the amount of digitized material is constantly growing. Typically, the digitized information is expressed without using common vocabularies for metadata annotations. As the metadata models and information content in the datasets are not harmonized, they are not directly interoperable, or able to communicate with each other, but instead the datasets form isolated silos.

The Web is a popular publication media for WW2 related information. However, this information is typically meant for human consumption only. The underlying *data* is not available in a machine-understandable, i.e., “semantic” format for research purposes and for end-user applications to utilize.

A fundamental problem with military historical data is making the contents mutually interoperable, so that they can be used and presented in a harmonized way [87]. *Semantic Web* technologies[1] provide solutions for combining heterogeneous isolated historical datasets [87, 137]. The key aspect of the success is the usage of vocabularies, ontologies, and existing classification systems [137]. A fundamental component of Semantic Web is the *Resource Description Framework (RDF)* [47], which is a data model and language for representing information using *Uniform Resource Identifiers (URIs)* or *Internationalized Resource Identifiers (IRIs)* to identify and describe resources. This way references to entities can be directed to the identity of the entity, instead of the entity name. This simple idea leads to a fundamental improvement in the interoperability of data and enables creating a more complete picture of the naturally very interlinked cultural heritage domain.

Data based on RDF is called *Linked Data* [23, 16] while the more global vision of an RDF-based distributed data graph is called the Semantic Web [18, 19, 184]. A Linked Data dataset is often called a *knowledge graph*, although a Linked Data dataset can also be considered a collection of RDF graphs [47]. Linked Data uses Semantic Web technologies that make it easily available on the Web, and understandable to both humans and machines [80].

Military history is a promising use case for Linked Data, as military historical data is by nature heterogeneous, distributed in various organizations, and expressed in different languages. Although the Semantic Web technologies are widely adopted in the whole cultural heritage domain [87], they have not been used much in the field of military history. Projects have created and published Linked Data about the domain, e.g., [216, 53, 152, 28], but this generally focuses on historical collection metadata, instead of actually representing the events and narratives of wars. In addition, the Linked Data projects have focused more on the First World War (WW1), instead of the more global, complex, and recent WW2, except in the domain of holocaust studies [15].

To make heterogeneous interlinked data usable for a wider audience, it is crucial to create user interfaces for the data that are easy to use. There are plenty of approaches to creating generic user interfaces for the Semantic Web [200, 129, 51, 17]. However, to best facilitate understanding and making sense of the data, the user interfaces should be adapted to the application domain. This has been a popular direction in the cultural heritage domain, where customized web portals are used to show different views to a knowledge graph for browsing, searching, analyzing, and visualizing different parts of the whole [190, 193, 117, 87].

As an interlinked dataset is updated and maintained, new kinds of challenges arise. The linking between the resources need to be kept in sync when changes as the contents are changed [8, 101, 204, 136, 169], i.e.,

[1] https://www.w3.org/standards/semanticweb/

handling change propagation within the dataset.

1.2 Objectives and Scope

The aim of this thesis is to improve the state of the art in representing and using military historical data as Linked Data. The goal is to also provide user interfaces to this data in a way that enables both conveying the information to interested “layman” users via web interfaces, while being also a useful resource to military history enthusiasts and scholars. The maintainability challenges introduced by interlinking heterogeneous data are discussed with a proposal for a solution.

The research contained in this thesis was conducted as a part of the WarSampo project[2], which integrates and publishes data concerning Finland in WW2 as *Linked Open Data (LOD)*. WarSampo is the first large scale system for serving and publishing WW2 LOD on the Web, published initially in 2015 [91].

The WarSampo infrastructure aims to support integrating new datasets into the knowledge graph in a sustainable way, by extending both the data model and data contents as needed. One hope of the project is that by making war data more accessible, the understanding of the reality of the war improves, which not only advances understanding of the past but also hopefully promotes peace in the future.

The research of this thesis concerns Finland in WW2, which defines the spatial and temporal boundaries of the used datasets. The used methods and Linked Data infrastructure are, however, meant to be applicable beyond these boundaries.

Figure 1.1 shows the main research areas of the thesis and summarizes the overarching research gap encompassing the combination of semantic reconciliation, semantic disambiguation, Linked Data maintenance, web portals, and military history as the applied domain. This thesis aims to fill the research gap currently existing between these research areas.

The aim of this thesis is to provide answers to the following research questions:

RQ1. How can wars be modeled and represented as data?

RQ2. How can heterogeneous military historical data be harmonized and integrated from distributed sources?

RQ3. How can military historical Linked Data be searched, browsed, analyzed, and visualized on web user interfaces?

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

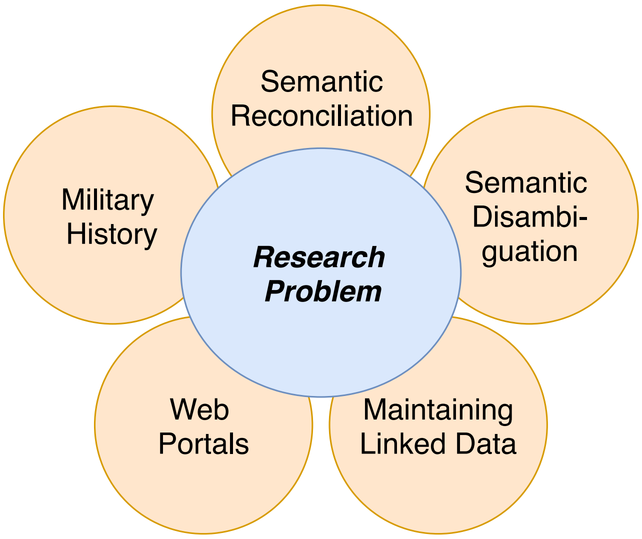


Figure 1.1. A diagram showing the research areas of the thesis, and depicting the combination of the research areas as the research gap of the thesis.

RQ4. How can the interlinked data and datasets be maintained?

The research questions are answered in the Publications I–IX. The connections between the research questions and research areas shown in Figure 1.1 are:

RQ1. Military History, Semantic Reconciliation
RQ2. Military History, Semantic Reconciliation, Semantic Disambiguation
RQ3. Military History, Web Portals
RQ4. Maintaining Linked Data

Table 1.1 shows how the publications are related to the research questions.

Publication	**RQ1.**	**RQ2.**	**RQ3.**	**RQ4.**
Publication I	**x**	**x**		**x**
Publication II	**x**			
Publication III	**x**			
Publication IV			**x**	
Publication V			**x**	
Publication VI			**x**	
Publication VII			**x**	
Publication VIII			**x**	
Publication IX				**x**

Table 1.1. The relationship between the publications and research questions.

1.3 Research Process and Dissertation Structure

Research in this thesis has been pursued using the *design science* [85, 161, 134, 72] research methodology. In contrast to natural sciences, which try to understand reality, design science attempts to create things that serve human purposes. This thesis does not attempt to give exhaustive answers to the research questions, but provides answers through applying design science to create various artifacts as part of the WarSampo project.

The outcomes of design science are useful artifacts, which can be *constructs*, *models*, *methods*, or *instantiations* [134, 85]. Constructs form the vocabulary of the domain, a model is a set of propositions or statements expressing relationships among constructs, a method consists of steps used to perform a task (e.g., an algorithm or a guideline), and an instantiation is a realization of an artifact in its environment [134]. The design science process consists of defining and motivating the problem, and then iteratively designing, developing, demonstrating, and evaluating the artifact using rigorous methods, and finally communicating the results to appropriate audiences [161].

The artifacts studied in the thesis are various parts of the WarSampo system. The artifacts that are both constructs and models are individual domain ontologies of the WarSampo ontology infrastructure. The WarSampo data model is a model artifact and the individual integrated datasets, as well as the whole knowledge graph and the semantic portal, are instantiations. Methods are designed and used to populate the WarSampo knowledge graph. The public WarSampo semantic portal acts as a proof of concept, demonstrating the suitability of the used artifacts for the purpose of representing and using military history as semantically harmonized data.

This thesis is structured as follows. In Chapter 2, the theoretical foundations are presented. In Chapter 3, the results of the publications are reviewed and summarized. Chapter 4 discusses the whole formed by the thesis, the significance of the results, the reliability and validity of the research, and provides suggestions for the directions of further research.

2. Theoretical Foundation

The research of this thesis builds on multiple research areas and research topics. In this section, the theoretical foundations and related work are presented in the research areas of the thesis: military history, semantic reconciliation, semantic disambiguation, web portals, and maintaining Linked Data.

2.1 Military History

History and Its Representation

History means the past as it is written, or orally transmitted, and the study thereof. Our connection to the historical past is built by historians through historical inquiry and interpretation [40, 223]. More generally, the concept of history, as the past, has an important role in human thought, framing the way we understand the reality [127].

According to [194], for most of the written history, narrative has been the main rhetorical device used by historians for historical writing. Narrative is defined as being a coherent story organized in a chronologically sequential order. The story is descriptive rather than analytical, and is concerned with people not abstract circumstances. Three types of relations can be observed between the events of a narrative for modeling purposes [138]: 1) temporal occurrence (the event occurs before, during, or after another event), 2) causality relation (the event is the cause or effect of another event), 3) mereological relation (the event is part of another event).

In addition to written history, history can also be directly approached via preserved tangible [203] and intangible [177] cultural heritage, which enable reinterpreting history, or to find more support for an existing interpretation.

Studying History

The intellectual tasks defining a historian's work are presented in [127]: 1) Historians strive to provide conceptualizations and factual descriptions of events and circumstances occurring in the past, answering questions like "what happened?" 2) Historians are often interested in answering "why" questions like "why did this event occur?" 3) Sometimes historians are interested in "how" questions like "how did this outcome come to pass?" 4) Often historians want to piece together the human meanings and intentions underlying a series of historical events.

A basic intellectual task of historians is discovering and making sense of the information stored in archives, which can be incomplete, ambiguous, contradictory, and confusing, requiring a great deal of interpretation [127]. A life cycle of historical information for historical research is presented in [27], consisting of six stages: 1) Creation, 2) Enrichment, 3) Editing, 4) Retrieval, 5) Analysis, and 6) Presentation.

Information sources and source criticism are essential in studying the past [223, 27, 138]. Many definitions exist of what actually is information, but all definitions agree that *information* is something more than *data* and something less than *knowledge* [27]. Knowledge is generally considered as a justified true belief [6], whereas information is often used without the requirement of truthfulness.

A disruption in historiography [14], the study of historical research, in the 20th century marked the rise of the "new history", or "new histories", which marks a shift of focus to employ statistics and methods of social science in research [70, 194].

Computationally oriented historical research, sometimes referred to as Historical Informatics [27], has been gaining momentum in the 21st century as part of the rise of the wider field of Digital Humanities [71, 36]. Harnessing the computational power readily available, and the growing collections of available data have proven fruitful for answering new kinds of questions about the past.

Prosopography, the study of collective biographies or groups of people [214, 205], has benefited from computational approaches that enable the gathering and analysis of large biographical datasets [34]. The idea is to analyze and compare groups of people based on their biographical information to find patterns and anomalies, and then try to explain the found phenomena.

Specificities of Military History

Much of the early written and oral history has discussed warfare and military history, e.g., Sun Tzu's *Art of War* in fifth century BC [46]. Wherever it has been studied, the purpose of military history has been to discover what actually happened in a war and why, and to transmit this information to soldiers, governments, and the people at large [46].

The scope of military history changed and broadened in the 20th century with the introduction of "new military history", which moved the field beyond narrow battlefield analysis towards studying the interface between war and society [44, 46], which has grown to be an integral part of modern military historical research [25, 21].

Military history can be considered part of cultural heritage, as wars and military always occur within a cultural context. Features of cultural heritage collection data are portrayed in [87]:

- **Multi-format**. The contents are in multiple formats, e.g., text documents, images, audio or video.

- **Multi-topical**. The contents concern various topics, e.g., art, history, artifacts.

- **Multi-lingual**. The contents are in different languages.

- **Multi-cultural**. The contents are related and interpreted in terms of different cultures.

- **Multi-targeted**. The contents are targeted to different user groups, e.g., laymen and domain researchers.

Military historical collection data shares all of the aforementioned features although being perhaps less varied in their topic. The topic variation increases if other related data sources are taken into account, that contain, e.g., personal information about the soldiers involved in a war, or art depicting a war. Multi-culturalism might not occur within a single country, but needs to be considered when combining data from multiple countries.

Key entities in the history of a war are the events of the military narrative, and the related entities, such as, people, military units, time, historical places [217]. Information on the key entities are depicted in various documents, such as photographs and person records. It is through these entities that an understanding about the whole of a war can be created. The entities are by nature interlinked, as the events usually involve people, either directly or through their military units, and happen at a certain place in a certain time. Photographs can document the people involved in an event and person records give their detailed personal information.

Military History of Finland

The military history of Finland as an independent state since 1917 contains different conflicts during both the First World War and the Second World War. Finland, as part of the Russian Empire before that, stayed mostly out of the WW1, and after the fall of the tsarist regime, declared its independence in 1917. Internal struggles and political polarization intensified, which led to the Finnish Civil War in 1918, resulting in the death of more than 38,000 people [197].

The history of Finland in the WW2 consists of three separate wars [123, 122, 111]: the Winter War, the Continuation War, and the Lapland War. The Winter War began with the Soviet Union invading Finland in November 1939 and ended in a peace treaty in March 1940. In the Continuation War from June 1941 to September 1944, Finland attacked the Soviet Union in an attempt to conquer back the areas lost in the Winter War, and was aided substantially by Germany in the process. In September 1944 Finland declared war to Germany, as part of the peace agreement with the Soviet Union, marking the start of the Lapland War, which continued until April 1945.

In recent years, the history of Finland in WW2 has still been an active research topic, with new themes for research emerging. Recent themes include e.g., psychological stress [104, 146], human-horse relationship [120], and a couple's relationship through letters [222], as well as global and national politics, holocaust, and various social aspects [103]. An archaeology project has recently been studying the cultural heritage of the German troops in Finland in the WW2 [111, 183]. A computationally oriented project has analyzed a large wartime photograph collection with data mining methods [60].

One of the main factors causing the political tensions that led to the Finnish civil war during the WW1 is considered the social inequality caused by the class system of the estates [173]. As studying the social aspects of war have become an important research area of historians in the 20th century, so have the facilities that support these approaches.

Occupational labels for people are commonly used in various person registers and these easily depict the approximate social class of a person, making them important resources in the study of social stratification and social mobility. The Historical International Standard of Classification of Occupations (HISCO) [208] is an important resource in studying these social aspects [67, 207, 116, 132]. In addition to HISCO, there is an existing Finnish classification [191].

Finnish fallen soldiers in WW2 were transported back to their hometown and buried in the so called "Heroes' Cemeteries" whenever possible [105, 122], making these cemeteries interesting for studying local histories.

A dataset of the people that died in the Finnish front in WW2, *Suomen* *sodissa 1939–1945 menehtyneet* (*Register of military deaths in the Finnish wars 1939–1945* in English [178]) is perhaps the most important existing dataset about Finland in the WW2. The creation of the dataset is summarized according to the description given by Lentilä [121]. The dataset was initially gathered as a register of the burial places of the people fallen in the wars. The foundations of the dataset are lists of people buried in the war memorial graveyards, originating from individual church parishes and gathered by the Finnish Church Council. The list of people was later supplemented with various data sources and additional information about the individuals was gathered from various sources. The dataset creation process started in 1985 and was still undergoing in 1997.

Another important resource is the online photograph archive, SA-kuva[1], of ca. 160,000 Finnish wartime photographs, portraying events at the war front, life on the home front, evacuation of Finnish Karelia, and other themes related to the war. Plenty of other information about Finland during the wars exist, but most of this is available only in a physical media in Finnish memory institutions, military history books, private collections, and so on. Some information is digitized, and more and more are being actively digitized. A large part of the digitized materials consist of handwritten documents, and although handwritten text recognition is an active research topic with improving results, the effort required to transform digitized hand-written material to text or data has been an important factor limiting the availability of cultural heritage data [78].

This thesis uses military history as the domain in which methods of computer science are developed and applied. The purpose is to create a harmonized view of the Finnish wars in WW2 as data, for the purpose of using the data in various applications that enable making sense of the data contents in intuitive ways. Furthermore, this thesis aims to bring military historical data more accessible to the wide public, and to support military historical research.

2.2 Semantic Reconciliation

To create an understanding about the complex realities of war, a fundamental task is to bring together information from various heterogeneous, distributed sources in an interoperable way. Semantic heterogeneity is a known obstacle to dataset integration, which can be solved by a process of *semantic reconciliation*, which makes the datasets interoperable with each other [65, 66, 90, 141].

Two levels of interoperability requirements can be observed in the reconciliation process [90]:

[1] `http://sa-kuva.fi/`

Syntactic interoperability. [115, 213] The same data can be structured in many ways on the syntactic level. The main step in achieving syntactic interoperability is formatting the data in the same format, such as using a shared database schema [160], XML schema [30, 115], or an RDF data model [159, 47]. In addition, data values can use various syntaxes. For example, dates are a typical example of this as they are represented differently in different countries, and their harmonization is needed using, e.g., XML Schema data types [165].

Semantic interoperability. [81, 77, 213] The meaning of different data fields can be different in different datasets. For example, dates can be syntactically interoperable, but given using different calendars, e.g., Julian or Gregorian. Also similar place names can have different meanings as there are many places globally with the same name. Such names need to be disambiguated to achieve semantic interoperability. People can also be referred to differently in different languages and depending on time. For example, a person may be referred to differently after getting married, or receiving a noble title or a position.

Linked Data

Despite historical popularity, relational databases are not considered ideal for managing heterogeneous and interlinked cultural heritage data [38], making it a promising use case for Linked Data [87]. RDF provides a flexible way to describe things in the real world, e.g., people and places, and how they are related to other things. The URIs used to identify Linked Data resources should be based on the *Hypertext Transfer Protocol (HTTP)*, so that information about them can be retrieved over the Web [16, 80]. URIs used with Linked Data should be persistent although in practice their reliability and persistency can be questioned as there is usually no authority to provide these in a trustworthy manner [13, 181, 39].

Numerous approaches have been proposed for creating Linked Data, for example, conversions from relational databases [84, 180], or using mappings, e.g., R2R [24], Karma [106], RML [55] or SPARQL Generate [118], which can ingest various source formats. An often used approach has been to program a custom pipeline [130, 150, 92, 167], which both converts source data into RDF and handles issues with semantic interoperability in the same process. A framework and tool for data fusion, conflict resolution, and quality assessment of Linked Data graphs is presented in [139].

In cases where it is important to track where the different pieces of information originate from, such as history, means to represent and encode such provenance information is required. There are various approaches to representing data provenance in Linked Data [228, 76, 227, 155].

The LOD Cloud[2] is a global endeavor to actualize the vision of the Semantic Web. It consists of Linked Data datasets that are linked to other datasets. Some of the important and most linked to datasets are DBpedia [7, 119], consisting of knowledge extracted from Wikipedia, Wikidata [162, 61], GeoNames, UK Governmental datasets [185], The Ontologies of Linguistic Annotations (OLiA), and WordNet.

Schemas and Ontologies

A metadata schema can be understood as "the semantic and structural definitions of metadata elements, including the relationships between those elements, which are represented in a standardized syntax or serialization format" [226]. An example of such schemas are database schemas, which provide the data model and structure of databases.

Ontologies are structured vocabularies that provide the semantics for entities and their relations, usually covering the concepts of a specific domain [73, 159]. RDF-based *RDF Schema*[3] and *OWL*[4] enable representing ontologies as an integrated part of the Linked Data. The ontologies can use different modeling frameworks, to best suit the modeling task at hand.

Ontologies that are based on simple, thesaurus-like structures are called *lightweight ontologies* [63, 68]. A common data model and vocabulary for expressing lightweight ontologies is the *Simple Knowledge Organization System (SKOS)*[5] [140, 10]. The focus in Semantic Web research has been shifting from the heavy use of formal semantics towards leveraging the collection of distributed, heterogeneous data using lightweight semantics [19].

Wars can be essentially seen as sequences of events, making the representation of events paramount in modeling military history. There are many approaches to modeling events in Linked Data [176, 170, 182, 186, 206]. The *CIDOC Conceptual Reference Model (CRM)*[6] is an event-based framework for modeling the heterogeneous domain of history, designed for the information exchange and integration of various cultural heritage metadata [57]. CIDOC CRM is an ISO standard (21127:2014), and widely used[7] [4] in the cultural heritage domain, e.g., in [158, 107, 150, 3, 92, 52]. There is a plethora of approaches to representing narratives as RDF [221, 138, 144, 50], of which many are compatible with CRM.

Other general-purpose metadata schemas in the cultural heritage domain

[2] `https://lod-cloud.net/`

[3] `https://www.w3.org/TR/rdf-schema/`

[4] `https://www.w3.org/TR/owl2-overview/`

[5] `https://www.w3.org/2009/08/skos-reference/skos.html`

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

[7] `http://www.cidoc-crm.org/useCasesPage`

are *Dublin Core (DC)* elements [8], its extended version *DCMI Metadata Terms (DCT)*[9] [4], and the *Europeana Data Model (EDM)*[10]. DC and DCT are created for representing essential document metadata elements in an interoperable way. The modeling rationales in EDM follow from the fact that it is mostly used for modeling tangible cultural heritage items like books, paintings, and films. The main rationales are 1) to distinguish between a cultural heritage object (item) and its digital representations, 2) distinguish between an item and its metadata record, 3) enable multiple metadata records for the same item, with possibly contradictory statements about the item. EDM is somewhat aligned with CRM [100, 163].

Military Historical Linked Data

There are several projects that have published Linked Data about the WW1, such as Europeana Collections 1914–1918[11], Collaborative European Digital Archival Research Infrastructure (CENDARI)[12] [28], Muninn[13] [216], Trenches to Triples[14] [32], Out of the Trenches [59], and WW1LOD [152].

Europeana is a digital platform for cultural heritage, publishing metadata of more than 53 million digitized cultural heritage objects from more than 3500 institutions. Europeana contains a smaller LOD pilot dataset that has metadata on ca. 2.40 million digitized texts, images, videos, and sounds [100]. EDM is used as the data model in Europeana.

The CENDARI project has identified six common types of entities interesting to historians studying both the medieval period and the WW1: places, people, institutions, dates, events, and topics [42]. The CENDARI project uses EDM as their data model, which is extended to the WW1 domain [42]. Information is integrated [28] from DBpedia, WW1LOD, 1914–1918-online, and Trenches to Triples.

The Muninn project has modeled historical military and civil organizations and their detailed structures, as RDF ontologies based on data from Wikipedia, with a focus on recording WW1 and facilitating data interchange in that domain [216]. The data model covers other related WW1 information, with stable ontologies[15] for military organizations, graves, and religions, and two related taxonomies of military terms.

Trenches to Triples project has created manual annotations of WW1

[8] https://www.dublincore.org/specifications/dublin-core/

[9] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/2012-06-14/?v=terms

[10] https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.8_102017.pdf

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

[12] http://www.cendari.eu/

[13] http://blog.muninn-project.org

[14] https://trenchestotriples.wordpress.com/

[15] http://rdf.muninn-project.org/

related catalogs held by the King's College [32]. The project sought to create a subject vocabulary of WW1 battles as LOD, which was used in indexing the catalogs.

Out of the Trenches has converted into RDF data about WW1 related war songs, war posters, newspaper articles, postcards, wartime records, soldier portraits and other archival materials, using a custom metadata model [59]. The data does not appear to be publicly available.

WW1LOD [152] uses a CIDOC CRM-based modeling of WW1 military history, containing information about events, actors, historical places, times, population statistics, keywords, and themes relating to the war. The main modeling rationales are: 1) Use existing established ontologies (especially CRM), 2) Model similar data uniformly, 3) Strive to model data intuitively, 4) Don't lose information included in the original source dataset. Events, actors, places, and times are modeled with mainly CRM. The population statistics are modeled with the W3C Data Cube vocabulary[16]. The notable deviations from CRM are 1) SKOS labels are used instead of CRM appellations, as there was no metadata about the actual appellations, 2) the relationships between organizations and people are modeled with CRM's shortcut property `crm:P107i_is_current_or_former_member_of` and additional relationships. These deviations are used to reduce the complexity of common data querying tasks. WW1LOD's geographical focus is on Belgium and the main purpose is to act as a reference vocabulary for other projects to link their WW1 collections to.

1914–1918-online[17] publishes historic and contemporary text articles about different aspects of WW1. It is based on Semantic MediaWiki, with RDF support and a simple metadata schema based on DCT and *Friend of a Friend (FOAF)* ontology[18].

There are a few works that use the Linked Data approach to WW2 or related holocaust studies.

The WW2 related narrative contents of the textual resources of the Bletchley Park Museum were modeled as Linked Data, using CRM, the Story and Narrative ontology [144], and a Bletchley Park domain ontology [45].

An important textual work in the Dutch WW2 historiography, "*Koninkrijk*", has been linked to structured SKOS metadata, and external resources [53].

The Jewish Contemporary Documentation Center (CDEC) has developed an online LOD database[19] on Italian holocaust victims and persecution events, and a related application for using the data [189]. They use custom classes and simple metadata annotations to describe resources [2]. The

[16] `https://www.w3.org/TR/vocab-data-cube/`

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

[18] `http://xmlns.com/foaf/spec/`

[19] `http://dati.cdec.it/lod/shoah/website/html`

dataset is part of the LOD Cloud[20] and contains "same as" links to other LOD Cloud datasets [31].

The Network for War Collections (Netwerk Oorlogsbronnen) initiative connects digitized collections about WW2 and holocaust in the Netherlands and publishes these as LOD in an Open Data Register[21] [210]. The collections are connected by manually mapping them to a WW2 thesaurus[22] of over 2300 concepts, containing links to the LOD Cloud.

The European Holocaust Research Infrastructure (EHRI) [2, 49] gathers and semantically integrates holocaust related databases, free text, and metadata. EHRI uses a set of controlled vocabularies for the metadata of heterogeneous cultural heritage resources of WW2 [210]. EHRI integrates metadata from distributed sources, employing technologies such as *Encoded Archival Descriptions*[23] and *Protocol for Metadata Harvesting*[24] [49], and has been experimenting with using Linked Data [54, 209], and have proposed to use CRM to represent people and events [2] and the Agent Relation Ontology AgRelOn [128] for modeling relations between people. The aforementioned Oorlogsbronnen and CDEC projects collaborate with EHRI.

This overview of the Linked Data based solutions for military historical information contains diverse projects that are mostly dealing with document metadata annotations. A few projects have striven to model the events of wars and the involvement of actors in them.

The data linking approaches used in the above projects are discussed in more detail in the next section.

A previous master's thesis has studied the representation of the Finnish military history narrative and wartime photographs in WarSampo [82]. This thesis aims to provide new understanding about how to achieve interoperability with heterogeneous datasets in the military history domain, through the various data integration and harmonization cases encountered in building WarSampo. In addition to new understanding, the idea is to create useful LOD ontologies and data for third parties to use, to promote interoperability in the future.

2.3 Semantic Disambiguation and Entity Linking

Semantic disambiguation is a key challenge in semantic interoperability [90], meaning the removal of uncertainty of meaning from possibly ambiguous textual representations or structured metadata. The main

[20] https://lod-cloud.net/dataset/shoah-victims-names

[21] https://www.oorlogsbronnen.nl/oorlogsbronnen-open-data

[22] https://data.niod.nl/WO2_Thesaurus.html

[23] http://www.loc.gov/ead/

[24] http://www.openarchives.org/OAI/openarchivesprotocol.html

problems are addressing synonymous and homonymous terms, e.g., identifying whether two place references with the same name actually refer to the same place or not. There is a plethora of approaches to this task [154, 26]. Knowledge is a fundamental component in word sense disambiguation [154], and structured external knowledge can be used in disambiguation problems in the form of ontologies. Usually, the disambiguation results do not need to be perfect for the resulting data to be useful, and the more effort is put into the disambiguation process, the better the results.

Named Entity Linking (NEL) (also *Entity Linking*, *Named Entity Disambiguation*) [187, 75, 35, 143] is the task of automatically disambiguating and linking the mentions of entity names in text to entities in a knowledge base. In the simplest form, the NEL process searches the text for the labels of knowledge base entities, and matches are linked. Linking accuracy can often be improved by employing, e.g., heuristics and candidate entity ranking [187, 143, 83, 212]. *Named Entity Recognition (NER)* [33] is the related problem of finding named entities of different categories without linking, typically from a text without a pre-existing knowledge base of entities.

A number of the created links are usually wrong, and some links missing, which is the trade-off for not having to go through the laborious process of manually creating the links. Several measures exist for evaluating the quality of the entity linking [187], of which *precision* is the fraction of correct links compared with all of the links:

$$\text{precision} = \frac{|\{\text{correctly linked entity mentions}\}|}{|\{\text{all generated links}\}|}$$

A related measure is *recall*, which is the fraction of the correctly linked entity mentions of all the entity mentions that should be linked:

$$\text{recall} = \frac{|\{\text{correctly linked entity mentions}\}|}{|\{\text{all entity mentions}\}|}$$

Finally, F_1 measure is the harmonic mean of precision and recall:

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Another related problem is that of finding matching structured data records between heterogeneous databases, called *record linkage* (also e.g., *data linkage*, *data matching*, *entity resolution*) [74, 43, 29, 98]. A typical scenario is matching people from two different person registers, which both contain structured data about each person expressed with different metadata schemas. A related problem is that of *duplicate detection* (also *deduplication*), in which the records are matched inside a database to find duplicates. The above defined precision and recall measures can be used also for record linkage [74].

Harmonizing and Linking Military Historical Data

Of the WW1 projects, the Europeana project relies on ingested collection metadata, which already conforms to its standards. In addition, a process is used to semantically enrich ingested metadata by linking them to GeoNames[25] for places, GEMET[26] for topics, Semium ontology[27] for time periods, and DBpedia[28] for people [100]. CENDARI extracts texts from documents, and uses NER and semantic disambiguation to create links to related information [28]. Muninn is based on using existing Wikipedia links [216]. WW1LOD harmonizes data from over ten different authoritative data sources. The data is converted into RDF and linked to vocabularies with a partly automated annotation process employing NEL, involving manual validation and correction [126, 152]. Some entities are manually indexed with keywords and themes [152].

Of the WW2 related projects, Koninkrijk [53] has linked the textual content of a historiographical text with two sources of structured knowledge. The text and a structured index of the text have been converted into RDF, after which another index has been created programmatically from the entities found when applying NER [33] to the structured text [53].

CDEC has created a domain ontology to express the information contained in heterogeneous databases [31]. The databases are transformed into RDF, and reconciled semantically in terms of the domain ontology, with a focus on holocaust victims and the persecution events.

The Oorlogsbronnen project connects collections by manually mapping them to a WW2 thesaurus [210]. EHRI integrates collection data, but does not use Linked Data for data reconciliation.

Quality measurements of the entity linking in the aforementioned projects are not publicly available.

This thesis aims to provide lessons learned in applying semantic disambiguating and entity linking in the military historical domain. The purpose is to be able to show what kind of approaches are expected to perform well in this domain.

2.4 Web Portals

Web portals are web sites that combine information from diverse sources in a uniform way[29]. The large public interest in military history can be

[25] http://www.geonames.org

[26] http://www.eionet.europa.eu/gemet/

[27] http://semium.org

[28] http://dbpedia.org

[29] https://en.wikipedia.org/wiki/Web_portal

observed in the plethora of web portals[30] dedicated to catering information in this field.

Sotapolku[31] is a military history portal, aiming at crowdsourcing the war paths of Finnish soldiers in WW2, as well as other information relating to them. The portal opened in December 2016, and employs a geospatial graphical user interface on which a military unit's path during the war is drawn, and the user can get more information about the individual steps. Sotapolku uses a relational database with a custom schema, which is also the case for many portals not employing Linked Data. They form silos that are not interoperable and do not communicate with other systems.

Conflict History[32] depicts more than 10,000 historical conflicts on a map, combined with a conflict timeline selector. The service was formerly a Flash application, visualizing data from the collaborative knowledge base Freebase [59], but is currently an iOS application.

There are also innovative approaches to user interfaces, such as the Fallen of World War II[33], which provides a data-driven documentary with interactive elements, using a variety of data sources. Our World in Data has published various visualizations of wars on the global scale with various time spans[34].

There are many ways of creating user interfaces for Linked Data, employing a variety of user interaction paradigms [23]. Additionally, the applications providing the user interfaces can be desktop applications, mobile applications, or Web based applications. *Rich Internet Applications* provide functionality like desktop applications, with the hypertext-based web's lightweight distribution architecture [62, 175].

As the Semantic Web technologies are based on the Web technologies, it is a natural choice to base Semantic Web application user interfaces on the Web technology stack.

Solutions are proposed to searching and browsing information on the Web of Data, based on text search, like Sig.ma [200], question answering, like PowerAqua [129] or FREyA [51], or browsing and link traversal, like Tabulator [17].

Semantic Portals

In addition to user interfaces for the whole Web of Data, there is a large variety of custom, localized domain specific applications, tailored to cater users with information needs of a specific domain. These *Semantic Portals*

[30] E.g., `https://ww2db.com/`, `http://www.world-war-2.info`, `https://www.britannica.com/event/World-War-II`

[31] `http://sotapolku.fi`

[32] `https://conflicthistory.com/`

[33] `http://www.fallen.io/ww2/`

[34] `https://ourworldindata.org/war-and-peace`

employ Semantic Web technologies to represent and harmonize information from multiple sources, and provide human access to the information via web user interfaces [190, 193, 117, 195].

Cultural heritage semantic portals [89] enable publishing complex interlinked data on web user interfaces, in which multiple semantic portal applications can create different perspectives to the whole dataset, based on, e.g., places, people, or other sub domain areas [95].

Useful search paradigms for the user interfaces of cultural heritage semantic portals include:

Geospatial search. [9, 1] The user can see resources on a map, and browse and explore them.

Temporal search. [138] Search based on a timeline component or a timespan selector.

Spatio-temporal search. [215] Geospatial search with a connected temporal search element.

Free text search. A free text search of the full metadata of the documents.

Field-restricted search.[35] A free text search, limited to certain field(s) of structured data [135, 188].

Faceted search. [201, 79, 157] The *Faceted search* paradigm (called also *view-based search* [168] or *dynamic hierarchies* [179]), is based on indexing data items along category hierarchies, i.e., facets (e.g., document types, places, etc.). The user can select categories on the facets in free order, and the data items included in the selected categories are considered the search results. After a selection, a count is calculated for each category, showing the number of results for that selection. The paradigm has been found especially suitable for Semantic Web user interfaces [86, 148].

Of the Linked Data based WW1 projects there are some that provide user interfaces to the Linked Data.

Europeana 1914–1918 is a WW1 related view[36] of the Europeana Collections portal [166], employing a faceted search interface.

CENDARI employs a faceted search of archival descriptions[37]. Additionally, there is a simple browser of the resources of several ontologies[38] [28].

[35] https://en.wikipedia.org/wiki/Full-text_search#Improved_querying_tools

[36] https://www.europeana.eu/portal/en/collections/world-war-I

[37] https://archives.cendari.dariah.eu/

[38] https://resources.cendari.dariah.eu/ontologies

WW1LOD is browsable and downloadable from the dataset homepage[39], and is integrated into a user interface for showing contextual information on an additional layer on top of text documents [152].

1914–1918-online[40] provides a user interface for browsing and searching historical and contemporary articles and pictures based on their Linked Data metadata.

A project about War Victims in Finland during the WW1 will publish Linked Data on a semantic portal [172].

The following user interfaces are used for WW2 related Linked Data. The Bletchley Park Museum Linked Data is usable via Bletchley Park Text application and a web page, intended to personalize museum experience and provide post-visit information [145].

The CDEC Digital Library is an online cultural heritage web portal displaying its own data together with data coming from the LOD cloud [31]. LOD Navigator[41] is an Electron[42] based JavaScript desktop application, providing an interactive spatio-temporal user interface [189] over the holocaust related events of 9040 people in the CDEC dataset. The user interface includes a geographical map and integrated timeline, with results shown as markers on the map, and a filter panel given to filter the result set, and facilitates quantitative and qualitative data analysis [189].

The Oorlogsbronnen data is presented on a Dutch cultural heritage portal[43] containing two different views to the collection metadata, i.e., the browsing of collections, and an upcoming view for searching people and displaying their biographical information [210].

EHRI Portal[44] contains descriptions of a large amount of holocaust related European archival institutions, individual archives, and authority sets on people and corporate bodies.

Visualizing Linked Data

Linked Data can be visualized in multiple ways, that can be divided into three classes [48, 102, 22]: 1) visualizing the graph structures, 2) visualizing data analysis results, like statistics, and 3) visualizing phenomena with different graphical methods, like viewing data on a map, on a time line, or using another suitable method.

Historical knowledge visualization in the Visual Contextualization of Digital Content (VICODI) project is discussed in [153]. The Narrative Building and Visualisation Tool uses an interactive timeline visualization as a key

[39] `http://www.ldf.fi/dataset/ww1lod/`

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

[41] `http://dh.fbk.eu/technologies/lod-navigator`

[42] `https://electron.atom.io/`

[43] `https://www.oorlogsbronnen.nl/`

[44] https://portal.ehri-project.eu

component in depicting narratives, complemented by graph visualizations and tables [138]. WikiStory enabled the browsing of Wikipedia-based biographies on a timeline [7]. BiographySampo provides various interactive views to Linked Data based biographies, including a spatio-temporal view of events, and visualization of social networks [94]. Troncy et al. [199] provide a spatio-temporal interface design for interactive visualizations of event-based Linked Data.

The aforementioned portals and visualizations provide different ways to show Linked Data on user interfaces in the cultural heritage domain. However, there is no single approach to displaying the complex history of a war on user interfaces. This thesis aims to develop new understanding about useful web user interfaces for military historical Linked Data. The purpose is to be able to create an understanding about wars beyond the possibilities of studying individual datasets traditionally.

2.5 Maintaining Linked Data

Maintaining ontologies is an important topic in facilitating interoperability on the Semantic Web, as the ontologies are rarely static, but instead are being adapted to changing requirements [131, 156, 224]. This *ontology evolution* raises new kinds of challenges.

Changes in an ontology may need to be propagated in three different scenarios: 1) inside the ontology, 2) to instances in a dataset using the ontology, and 3) to depending ontologies and applications [192]. A variety of infrastructures [131, 64] and tools [225, 164] have been proposed for handling ontology evolution. The same challenge of change propagation is evident with all of linked open data [8, 101, 204, 136, 169], often called *Linked Data Dynamics* in this context. Research in this field is concerned with detecting, describing, and propagating changes, as well as versioning of Linked Open Data resources and datasets.

Database-schema evolution [156, 133, 5, 11] is related to ontology evolution and ontology evolution research builds on the prior research of that field. Schema integration [12, 69] is a related problem faced when combining heterogeneous databases, leading [56] to the challenges of semantic reconciliation.

Data quality is an important topic relating to data maintenance. Linked Data validation measures structural data quality of the Linked Data graphs, using some kind of explicit definition of the expected structure [113]. Validation improves the reuse potential of the data and ontologies. The two modern approaches for Linked Data validation are *Shape Expressions (ShEx)* and *Shapes Constraint Language (SHACL)* [113]. A master's thesis studying RDF validation has recently validated the events graph of the

WarSampo knowledge graph using SHACL [112]. According to the results of the validation the data is deemed quite valid, with 38 reported violations in the 20,000 triples, of which 28 violations are empty strings used as property values. The validation uses 6 rather simple constraints, which do not make good use of the CIDOC CRM structures as the validation is restricted on the event data without other related information.

This thesis seeks to provide new understanding of the different change propagation scenarios faced when maintaining a set of interlinked graphs about military history and to provide methods to handle the change propagation scenarios in practice.

3. Results

The following presents answers to the research questions of the thesis in detail. The results are compared against the current state of the art. The results as a whole are further reflected against previous research in Chapter 4.

3.1 Modeling and Representing Military History

The research question 1 concerns modeling military historical information as data.

RQ1. How can wars be modeled and represented as data?

Publications I–III provide solutions to this question by presenting the data modeling rationales employed in the WarSampo project. The focus is on modeling tangible and intangible cultural heritage relating to Finland in WW2. This information includes primary and secondary sources, as well as interpretations made by historians.

State of the Art

The state of the art in modeling military history as data is to use Linked Data and base the metadata schema on either CIDOC CRM or EDM. The focus of CRM is on harmonizing cultural heritage metadata with event-based fine-grained modeling. The focus of EDM is on harmonizing metadata about diverse cultural heritage collections using object-centric, event-centric, or both modeling approaches [99]. Both of these models facilitate interoperability in harmonizing datasets, and the choice of the metadata schema is mostly an opinion, concerning whether one wants to harmonize information about the actual historical events (CRM), or harmonize cultural heritage collection or object metadata (EDM).

Of the state-of-the-art projects, CRM is employed in WW1LOD [152].

In addition, CRM was used in modeling the Bletchley Park Museum resources [45] and EHRI has proposed to use CRM to represent people and events [2]. EDM is used in Europeana Collections and CENDARI [28]. Simpler metadata schemas, such as DCT, are used in Muninn [216], CDEC [2], Oorlogsbronnen [210], and EHRI [210], Koninkrijk [53].

Improving on the State of the Art

Publication I answers the question by providing the data model used in WarSampo for representing military history, and the source datasets.

As several heterogeneous, distributed sources make references to the same key entities, it is crucial that there is a standard way of referring to the entities when combining this information. Linked Data enables this, as URIs create an identity to each key entity. Furthermore, the different relations between entities can be separated by different properties with formal semantics. URIs enable the linking of pieces of information directly to the key entities in a sustainable way, as anyone can re-use the same identifiers.

Wars can be essentially seen as sequences of events, supporting the use of CIDOC CRM as the conceptual framework for representing event-based historical information from heterogeneous data sources, and enabling to create a unified view of a military history narrative as a sequence of events that can be placed on a timeline. The actions and events of involving actors, such as the wounding, promotion, or death of a soldier, and the formation or movement of a military unit can be described naturally as events.

The WarSampo data model builds on the state-of-the-art data model employed in WW1LOD: using CRM as the backbone with minor documented deviations, and using a few other useful ontologies for metadata annotations. The state of the art is improved in WarSampo by extending CIDOC CRM for the military history domain and especially for the representation of the events of war, by creating RDFS subclasses of CRM classes. This enables semantically separating the classes, e.g., for information retrieval purposes. This approach of creating specialized subclasses is endorsed in EDM [99], but not explicitly supported in CRM. However, the RDF data model enables this when using CRM as RDF. For example, the activity of a person can be further divided into military activity or photography, of which the former can be grained down to battles, bombardments, and promotions. This enables to easily select and examine, e.g., battles as a separate activity. The same approach can be used for properties, to use more specific subproperties of those existing in CRM.

The CRM extensions are created based on the need to represent information in the source datasets, i.e., a new subclass is only created should there be instances of the class created based on the source datasets. The variety of the WarSampo source datasets is enough to cover the military historical key entities in the data model, but the scope of the data model does not cover everything relating to military history. However, the data model can be easily extended as needed, by creating new specialized RDFS subclasses of either the CRM classes or the already specialized classes. For example, one could create a new class for a specific type of military activity, such as aerial combat, by creating a new subclass of the existing military activity class in the WarSampo data model. Similarly new specialized properties can be created based on the ones existing in CRM or in the WarSampo data model by defining a new property as RDFS subproperty of an existing one. For example, a new subproperty of the CRM property `crm:P12_occurred_in_the_presence_of` could be created for depicting the involvement of an aircraft in an aerial combat. A new class for aircrafts could be defined as a subclass of CRM `crm:E24_Physical_Man-Made_Thing`.

Table 3.1 presents the source datasets of WarSampo. The source datasets are provided by various organizations, such as the National Archives of Finland, The Finnish Defence Forces, The Association for Military History in Finland, The National Land Survey of Finland, and the Aalto University. The source datasets were in different formats, e.g., spreadsheets, text, web pages, images, *application programming interfaces (API)*, *Extensible Markup Language (XML)* documents, *Portable Document Format (PDF)* documents, and RDF graphs. The contents of the source datasets did not use any shared vocabularies or shared practices of referring to entities.

As a basis for harmonizing the heterogeneous source datasets, an ontology infrastructure was first constructed from some of the source datasets, to which the other datasets can be linked to. The ontology infrastructure consists of domain ontologies (DO) modeled using 1) CIDOC CRM: people, military units, places, and military ranks, and 2) SKOS: citizenships, genders, marital statuses, mother tongues, nationalities, perishing categories, and occupations.

For the created WarSampo RDF resources, the source dataset where the resource is originating from is generally annotated directly to the resource. In the case of the Prisoners of War dataset, the source data contains also detailed information about original information sources for individual pieces of information, and often multiple contradicting values for a single spreadsheet column. The detailed information sources are modeled in WarSampo as RDF Reifications [227] with a source annotation. Reifications also enable the ranking of multiple values for a single property or attaching various annotations, like other provenance information, to any RDF triples.

Descriptions of how key entities, i.e., events, places, documents, people, military units, and occupations, are modeled are given next.

Events. WarSampo events have been classified into 19 subclasses of the class `crm:E5_Event`. They are used to model war and political events like battles, bombardments, or political activity, and events of the actors partic-

Table 3.1. The source datasets of WarSampo.

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

ipating in the war, like births, joinings to military units, troop movements, dissolutions, and promotions. Each event has a textual representation, a time-span, links to participating actors, and information where the event occurred with links to the place ontologies when applicable.

Photography events are created for photographs to 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.

Places. The ontology of places is combined from four different sources, and modeled with a simple schema, which contains properties for the place name, coordinates, polygon, place type, and part-of relationship of the place. Each place is an instance of a subclass of `crm:E53_Place`.

Documents. War related document files, i.e., photographs, war diaries, and magazine articles, are modeled as separate subclasses of `crm:E31_Document`, having Dublin Core like metadata annotations. A separate group of documents are person records, i.e., death records and war prisoner records, which are linked to corresponding person instances via `crm:P70_documents` relations. Person records are directly linked to the ontology infrastructure.

Actors. Publication II presents a key part of the WarSampo ontology infrastructure: the actor ontology, consisting of people and military units. Contrary to actor vocabularies, the actor ontology represents an actor as a biographical life story. The `crm:E39_Actor` class, with its subclasses can be seen to be central to the whole data model, as there is a considerable amount of references to them from the other classes. Actors are modeled mostly using CRM, by re-interpreting the information in the source datasets as events relating to the actor. For people, CIDOC CRM-based *Bio CRM* [202] is used to present roles like occupations.

The military units are particularly challenging: the army hierarchy is large and changes rapidly, unit identification codes and names change occasionally to confuse the enemy, and casualties and replacements constantly change unit compositions. The army hierarchy, including the temporal changes made in it, is modeled as the events of a unit joining its superior unit.

Occupations. Publication III presents the creation of the SKOS-based domain ontology of Finnish historical occupations, AMMO, which is based on thousands of Finnish historical occupational labels from the early 20th century, also containing the occupational labels of WarSampo. It improves the state of the art by combining synonymous occupational labels into single concepts containing multiple labels. This greatly enhances the studying of the people in the WarSampo data through their occupations. AMMO provides a resource for the prosopographical study of the person registers, as most of the people in them are annotated with occupational labels. AMMO is aligned with the international HISCO standard and the

Finnish Classification of Occupations to provide social stratification information and field of work information, and for international and national interoperability. Domain ontologies such as AMMO can be used as natural components for faceted search and semantic recommendation in semantic portals for military history. AMMO is to the best of our knowledge, the first Finnish occupation ontology.

3.2 Harmonizing Heterogeneous Data

The research question 2 concerns attaching data and documents to the representational model of military history devised in research question 1. Heterogeneous data need to be combined from various sources, and in various formats, to create a unified, interoperable whole of the history of Finland in WW2. The contributions are considered on the three levels of semantic reconciliation [90]: 1) syntactic interoperability, 2) semantic interoperability, and 3) semantic disambiguation.

RQ2. How can heterogeneous military historical data be harmonized and integrated from distributed sources?

The Publication I provides answers to this question by presenting the methods and implementation of the integration and harmonization of heterogeneous datasets about Finland in WW2 into the WarSampo Linked Data infrastructure. The state of the art is first presented for comparison.

State of the Art

The state of the art in providing syntactic interoperability is to use the RDF framework of Linked Data, used in all of the referenced state-of-the-art projects, except EHRI [210]. The method of achieving semantic interoperability is by using shared metadata schemas, such as CRM and EDM, and shared ontologies, such as DCT and FOAF. This method can be implemented by manually annotating resources, or by using NEL.

The method for achieving semantic disambiguation is by resolving the identities of entities contained in text or structured data, typically referred to by entity names. The scope of the problem is to identify the identities within a dataset or project, so that two mentions of an entity, e.g., a person, refer to the same identity. Semantic disambiguation has been implemented by manually annotating resources by domain experts, or using NEL, or both. Record linkage does not seem to have been used in previous research in the military history domain.

Improving on the State of the Art

Publication I presents a method for harmonizing and integrating heterogeneous, distributed datasets into a common data model. The method is then applied to create the WarSampo knowledge graph, by populating the data model from the source datasets.

The method uses Linked Data to provide syntactic interoperability. By using CRM and a shared ontology infrastructure, the heterogeneous source datasets can be reconciled semantically to refer to shared entities instead of referring to entities by names. E.g., instead of referring to a person by their name, the person can be referred to by a URI, to make an unambiguous reference to a certain person. The resulting data graphs, which use the DOs, are referred to as *metadatasets (MDS)*.

Various data transformation processes can be used to transform datasets into the harmonizing data model, and link entities in the created MDSs to the DOs. Information like military ranks, places, and occupations is typically given as text strings in the source datasets. Linking these to DOs is usually rather simple, by comparing the text strings with the labels of resources in the DOs, but to improve recall, some programmatic harmonization and heuristics have been used. The linking enables information retrieval based on the DOs. For example, by linking entities to places, it is easy to retrieve all information relating to a place, from several datasets.

Information about a person can be found in various datasets, each bringing some new information about the person, which can be used to create a more and more full biography of a single person. However, the challenge is that the person can be referred to very differently in different data, as e.g., the military rank and military unit of a soldier can change in time, and often the name of a person is not given in full. Details may be missing, like the date of birth, or they may even be incorrect. The same full name can refer to different people, and different names can refer to the same person, as people have changed names. In the early 20th century it was common to take a new Finnish surname to replace a former non-Finnish one.

The entity linking enriches the MDSs. For the entities in an MDS, the DO may contain further information about the linked concept, such as in the case of AMMO occupations, where linking a person to an occupation concept also enriches the person with information about his social status through the occupation. The DOs can also be used to provide contextual information through the entity linking, e.g., people with the same occupation.

The process of populating the data model to create the WarSampo knowledge graph started by creating the shared DOs. The source datasets were then converted into RDF and linked to the DOs to create the WarSampo MDSs. Some early DOs, i.e., 5610 people, military units, military ranks, and medals, involved manual ontology editing, and the processes used to create them are not repeatable. They are maintained directly in RDF format, along with the SKOS-based DOs used by the person records, i.e., citizenships, genders, marital statuses, mother tongues, nationalities, and perishing categories.

A repeatable data transformation pipeline is used for building the majority of the knowledge graph 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. In this method, the domain experts can maintain the primary data in the original native format. When a source dataset is updated, the pipeline can be used to easily recreate the whole knowledge graph with the updates.

The semantic disambiguation is mostly implemented using different NEL [75] implementations, e.g., [149, 196], to link resources to the DOs. In the linking, 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 providing 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 user's ability to evaluate the links and give feedback if a link is clearly wrong.

When NEL is used to link textual terms 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.

In some cases, like when disambiguating person records in different datasets, more emphasis needs to be put on precision. The person records are matched to already existing person instances using probabilistic record linkage [74], with a logistic regression-based machine learning implementation. New person instances are created in the Persons DO for the person records that don't match any existing person.

The resulting WarSampo knowledge graph [109] consists of 14,300,000 triples. The core classes used in both MDSs and DOs are presented in Figure 3.1, with instance counts and main linkage between the class instances. The arrow direction depicts the direction of linking and LOD Cloud refers to the global LOD cloud. The core classes contained within a DO are shown as green rectangles and the MDSs using the DOs are shown with yellow rounded rectangles.

The NEL of war and political 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 [83]. 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 [83]. 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 [83].

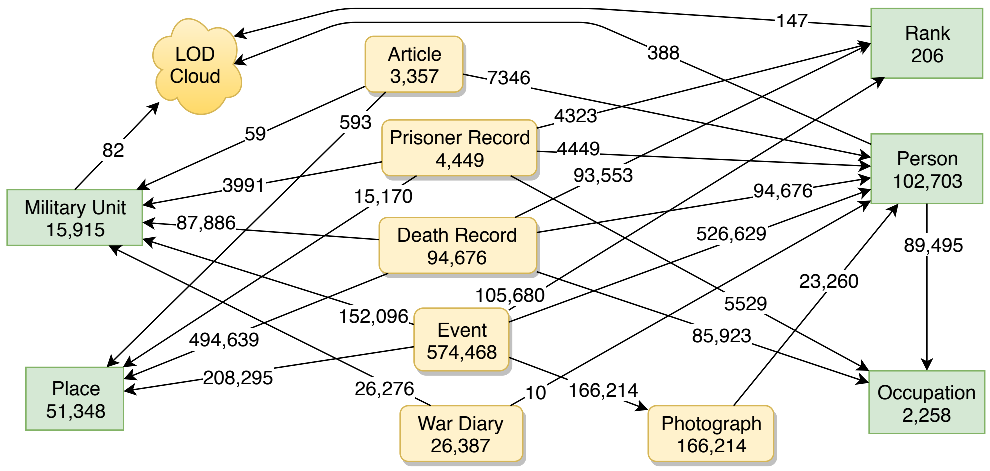


Figure 3.1. The core classes with instance counts and linkage between class instances.

The person record linkage of death records results in 613 death records linked to matching people in the 5611 pre-existing person instances, while for the remaining 94,056 death records, new person instances are created.

The person record linkage of prisoner records results in 1397 person records linked to matching people in the 99,667 pre-existing person instances, while creating 3031 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 1397 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.

In addition to new understanding about the applicable methods, the created artifacts, i.e., the data model, DOs, and MDSs facilitate interoperability themselves. They are available for anyone to use and link to, helping to prevent future interoperability problems.

3.3 Semantic Portal For Military History

Research question 3 deals with using the Linked Data to make sense of military history via web-based user interfaces.

RQ3. How can military historical Linked Data be searched, browsed, analyzed, and visualized on web user interfaces?

The publications IV–VIII provide answers to this question by presenting the WarSampo portal and its different perspectives to the interlinked military historical data. The state of the art is first presented for comparison.

There seems to be only one existing online system for providing user interfaces for Linked Data about historical war events: The LOD Navigator uses a spatio-temporal interactive user interface of the holocaust related events of 9040 people in the CDEC dataset [189]. Many of the systems for WW1 or WW2 related information employ basic browsing and searching functionality of the metadata of collections and tangible cultural heritage items.

As a basis for the design of the user interfaces, Publication V lays out the different user groups of military historical data identified at the National Archives of Finland. The military historical data users can roughly be divided into three groups: 1) academic researchers, 2) military history enthusiasts, and 3) private citizens.

The first group has the widest range of needs regarding the data, but often has the best skills to handle and refine the data by themselves. The focus of academic researchers seems to be shifting from a macro level towards studying individuals and the social aspects of war [25, 21].

Military history enthusiasts usually approach the data from a military unit perspective, or they may concentrate on a certain location during a narrow time frame. They may also be searching for irregularities, such as peaks in the numbers of casualties or in certain age groups within the data.

Private citizens usually begin their search for information with their own relatives who were lost during the war. After finding that out, they may go on searching for similar destinies based on age group, unit, or locations (e.g., home towns or the location where their relatives lost their lives). Private citizens are usually the most dependent on easy-to-use user interfaces. It seems apparent that this is the largest user group of the data.

WarSampo is targeted to all of the three user groups. The WarSampo dataset can be queried directly from the open SPARQL endpoint[1], or downloaded and processed further, by e.g., academic researchers. The WarSampo portal provides user-friendly applications for all the user groups to search, browse, analyze, and visualize the data.

Publication IV introduces the WarSampo portal[2], which is a semantic portal improving the state of the art by providing nine different perspectives on the WW2 related Linked Data. Event-based spatio-temporal user interfaces enable depicting the whole war as events, completed with perspectives for searching and browsing related resources like people, places and photographs. The high level of interlinking within the knowledge graph is used to provide links between resources in different perspectives.

The perspectives are a collection of interlinked applications, which ad-

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

[2] `https://www.sotasampo.fi/en/`

dress different end-user information needs. 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. The different perspectives are supported without modifying the data, but by only adjusting the data queries sent to the open SPARQL endpoint of the LOD service. In this way new application perspectives to the data can be added easily and independently, without affecting the other perspectives. The list of perspectives is given in Table 3.2.

Perspective	**Search Paradigms**	**Results Display**
Events	spatio-temporal	spatio-temporal, event home page
Persons *Publication VI*	free text search	person home page
Military Units	free text search	spatio-temporal, military unit home page
Places	geospatial, free text search	geospatial, home page
Articles	faceted search	table, contextual reader
Casualties *Publication VIII*	faceted search	table, visualizations
Photographs *Publication VII*	faceted search	table, photograph home page
War Cemeteries *Publication V*	faceted search	table, cemetery home page
Prisoners *Publication VI*	faceted search	table, visualizations

Table 3.2. The WarSampo application perspectives, referenced to Publication IV unless otherwise stated.

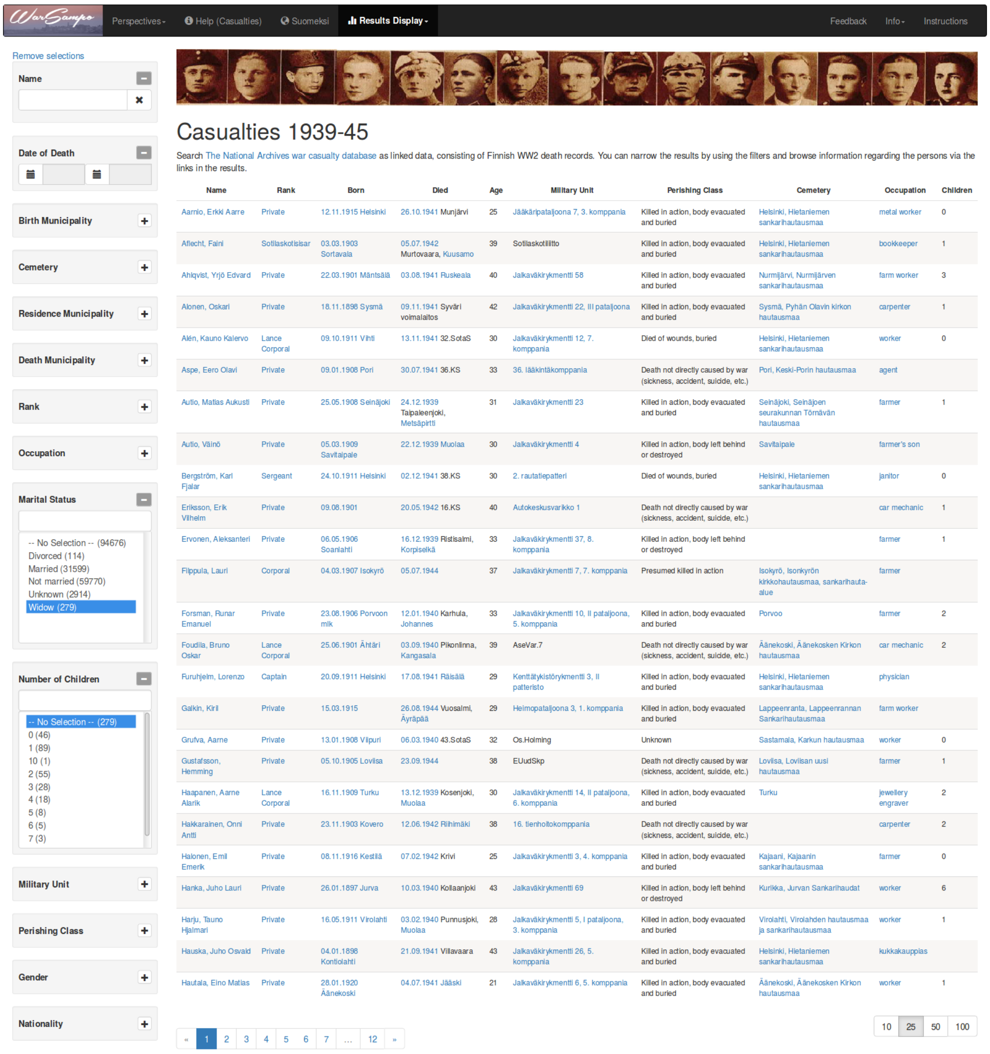
All perspectives employ a search over the entities of interest in the particular perspective. The used search paradigms are:

Geospatial search. The user can search visually on a pannable and zoomable map, overlayed with markers or polygons highlighting the places containing results. The user can select the place of interest, to see either the home page of a resource, or links to resources related to the place, depending on the perspective.

Spatio-temporal search. Same as above, with the addition of an interactive, visual timeline component, on which the relevant events are shown on the date of their occurrence. The user can change the focused time period by scrolling the timeline horizontally.

Free text search. The user can search resources by entering a part of the name into a text input field.

Faceted search. The user can interactively explore, browse, and analyze the resources. Faceted search is based on displaying categories for each facet, from which the user can select one, which then narrow down the result set to include only the results that match the user selections. Facets are presented on the left of the user interface with free text search support. The number of hits on each facet is calculated dynamically and shown to the user, and selections leading to an empty result set are hidden. The faceted search is used not only for searching but also as a flexible tool for researching the underlying data. The faceted search of the casualties perspective is shown in Fig. 3.2, where the hit counts immediately show distributions of the result set along the facet categories.



Marital Status:
-- No Selection -- (94676)
Divorced (114)
Married (31599)
Not married (59770)
Unknown (2914)
Widow (279)

Number of Children:
-- No Selection -- (279)
0 (46)
1 (89)
10 (1)
2 (55)
3 (28)
4 (18)
5 (8)
6 (5)
7 (3)

Casualties 1939-45

Search The National Archives war casualty database as linked data, consisting of Finnish WW2 death records. You can narrow the results by using the filters and browse information regarding the persons via the links in the results.

Name	Rank	Born	Died	Age	Military Unit	Perishing Class	Cemetery	Occupation	Children
Aarnio, Erkki Aarre	Private	12.11.1915 Helsinki	26.10.1941 Munjärvi	25	Jääkäripataljoona 7, 3. komppania	Killed in action, body evacuated and buried	Helsinki, Hietaniemen sankarihautausmaa	metal worker	0
Aflecht, Faini	Sotilaskotisisar	03.03.1903 Sortavala	05.07.1942 Murtovaara, Kuusamo	39	Sotilaskotiliitto	Killed in action, body evacuated and buried	Helsinki, Hietaniemen sankarihautausmaa	bookkeeper	1
Ahlqvist, Yrjö Edvard	Private	22.03.1901 Mäntsälä	03.08.1941 Ruskeala	40	Jalkaväkirykmentti 58	Killed in action, body evacuated and buried	Nurmijärvi, Nurmijärven sankarihautausmaa	farm worker	3
Alonen, Oskari	Private	18.11.1898 Sysmä	09.11.1941 Syväri voimalaitos	42	Jalkaväkirykmentti 22, III pataljoona	Killed in action, body evacuated and buried	Sysmä, Pyhän Olavin kirkon hautausmaa	carpenter	1
Alén, Kauno Kalervo	Lance Corporal	09.10.1911 Vihti	13.11.1941 32.SotaS	30	Jalkaväkirykmentti 12, 7. komppania	Died of wounds, buried	Helsinki, Hietaniemen sankarihautausmaa	worker	0
Aspe, Eero Olavi	Private	09.01.1908 Pori	30.07.1941 36.KS	33	36. lääkintäkomppania	Death not directly caused by war (sickness, accident, suicide, etc.)	Pori, Keski-Porin hautausmaa	agent	
Autio, Matias Aukusti	Private	25.05.1908 Seinäjoki	24.12.1939 Taipaleenjoki, Metsäpirtti	31	Jalkaväkirykmentti 23	Killed in action, body evacuated and buried	Seinäjoki, Seinäjoen seurakunnan Törnävän hautausmaa	farmer	1
Autio, Väinö	Private	05.03.1909 Savitaipale	22.12.1939 Muolaa	30	Jalkaväkirykmentti 4	Killed in action, body left behind or destroyed	Savitaipale	farmer's son	
Bergström, Karl Fjalar	Sergeant	24.10.1911 Helsinki	02.12.1941 38.KS	30	2. rautatiepatteri	Died of wounds, buried	Helsinki, Hietaniemen sankarihautausmaa	janitor	0
Eriksson, Erik Vilhelm	Private	09.08.1901	20.05.1942 16.KS	40	Autokeskusvarikko 1	Death not directly caused by war (sickness, accident, suicide, etc.)		car mechanic	1
Ervonen, Aleksanteri	Private	06.05.1906 Soanlahti	16.12.1939 Ristisalmi, Korpiselkä	33	Jalkaväkirykmentti 37, 8. komppania	Killed in action, body left behind or destroyed		farmer	1
Filppula, Lauri	Corporal	04.03.1907 Isokyrö	05.07.1944	37	Jalkaväkirykmentti 7, 7. komppania	Presumed killed in action	Isokyrö, Isonkyrön kirkkohautausmaa, sankarihauta-alue	farmer	
Forsman, Runar Emanuel	Private	23.08.1906 Porvoon mlk	12.01.1940 Karhula, Johannes	33	Jalkaväkirykmentti 10, II pataljoona, 5. komppania	Killed in action, body evacuated and buried	Porvoo	farmer	2
Foudila, Bruno Oskar	Lance Corporal	25.06.1901 Ähtäri	03.09.1940 Pikonlinna, Kangasala	39	AseVar.7	Death not directly caused by war (sickness, accident, suicide, etc.)	Äänekoski, Äänekosken Kirkon hautausmaa	car mechanic	2
Furuhjelm, Lorenzo	Captain	20.09.1911 Helsinki	17.08.1941 Räisälä	29	Kenttätykistörykmentti 3, II patteristo	Killed in action, body evacuated and buried	Helsinki, Hietaniemen sankarihautausmaa	physician	
Galkin, Kiril	Private	15.03.1915	26.08.1944 Vuosalmi, Äyräpää	29	Heimopataljoona 3, 1. komppania	Killed in action, body evacuated and buried	Lappeenranta, Lappeenrannan Sankarihautausmaa	farm worker	
Grufva, Aarne	Private	13.01.1908 Viipuri	06.03.1940 43.SotaS	32	Os.Holming	Unknown	Sastamala, Karkun hautausmaa	worker	0
Gustafsson, Hemming	Private	05.10.1905 Loviisa	23.09.1944	38	EUudSkp	Death not directly caused by war (sickness, accident, suicide, etc.)	Loviisa, Loviisan uusi hautausmaa	farmer	1
Haapanen, Aarne Alarik	Lance Corporal	16.11.1909 Turku	13.12.1939 Kosenjoki, Muolaa	30	Jalkaväkirykmentti 14, II pataljoona, 6. komppania	Killed in action, body evacuated and buried	Turku	jewellery engraver	2
Hakkarainen, Onni Antti	Private	23.11.1903 Kovero	12.06.1942 Riihimäki	38	16. tienhoitokomppania	Death not directly caused by war (sickness, accident, suicide, etc.)		carpenter	2
Halonen, Emil Emerik	Private	08.11.1916 Kestilä	07.02.1942 Krivi	25	Jalkaväkirykmentti 3, 4. komppania	Killed in action, body evacuated and buried	Kajaani, Kajaanin sankarihautausmaa	farmer	0
Hanka, Juho Lauri	Private	26.01.1897 Jurva	10.03.1940 Kollaanjoki	43	Jalkaväkirykmentti 69	Killed in action, body left behind or destroyed	Kurikka, Jurvan Sankarihaudat	worker	6
Harju, Tauno Hjalmari	Private	16.05.1911 Virolahti	03.02.1940 Punnusjoki, Muolaa	28	Jalkaväkirykmentti 5, I pataljoona, 3. komppania	Killed in action, body evacuated and buried	Virolahti, Virolahden hautausmaa ja sankarihautausmaa	worker	1
Hauska, Juho Osvald	Private	04.01.1898 Kontiolahti	21.09.1941 Villavaara	43	Jalkaväkirykmentti 26, 5. komppania	Killed in action, body evacuated and buried	Helsinki, Hietaniemen sankarihautausmaa	kukkakauppias	
Hautala, Eino Matias	Private	28.01.1920 Äänekoski	04.07.1941 Jääski	21	Jalkaväkirykmentti 6, 5. komppania	Killed in action, body evacuated and buried	Äänekoski, Äänekosken Kirkon hautausmaa	worker	1

Figure 3.2. The faceted search interface of the casualties perspective with one selected facet. The left side contains the facets, displaying available categories and the amount of death records for each selection category. Death records matching the current facet selections are shown as a table.

There are many different approaches to displaying the results based on the search paradigm. Many perspectives combine multiple types of results display, either as complementary parts of the perspective or as optional ways to display a certain result set. The used results display types are:

Geospatial. The results are shown visually on a map as markers or polygons, which can be clicked to show the home page of the place or event, depending on the perspective. The user can choose to view the results on top of digitized historical maps.

Spatio-temporal. Similar as above, but with an additional timeline element. A heatmap overlay shows casualties on the map during the selected time-frame.

Table. The typical results display of a faceted search perspective, listing the results with their most important details.

Visualizations. The results of faceted search can be visualized based on the properties of the result class, to study the distributions of values in that result set. Publication V presents prosopographical visualizations based on the death records. Visualizations in the casualties perspective include various bar and chart visualizations as alternative results displays to the table view. Figure 3.3 presents a screenshot of the novel sankey diagram of soldier life paths, showing the life paths of the 40 soldiers buried in the cemetery of Inari in Ivalo. The diagram shows where the soldiers were born, where they lived, where they died, and where they are buried. Additionally, the war cemetery home page visualizes prosopographical statistics of the buried people.

Contextual reader. The articles perspective, presented in Publication IV, uses a faceted search of the Kansa Taisteli magazine articles. The articles can be read with an overlay providing contextual information, with real-time annotations based on EL [151]. The annotations link to WarSampo DOs and DBpedia, and work as hyperlinks to further information.

Home page. Of the nine perspectives, five provide home page for the contained entities, by employing a systematic URI referencing policy. The home pages are domain specific *Hypertext Markup Language (HTML)* pages for human usage. For example, a soldier in the "persons" perspective, has a home page, created by the perspective, that can be linked easily to the home pages of the other perspectives by their URIs. All of the home pages contain links to related photographs, people, and military units. Also home pages of four entity types link to events, and three to magazine articles. Non-personalized semantic recommender systems [147] based on entity linking are used to provide links for further information.

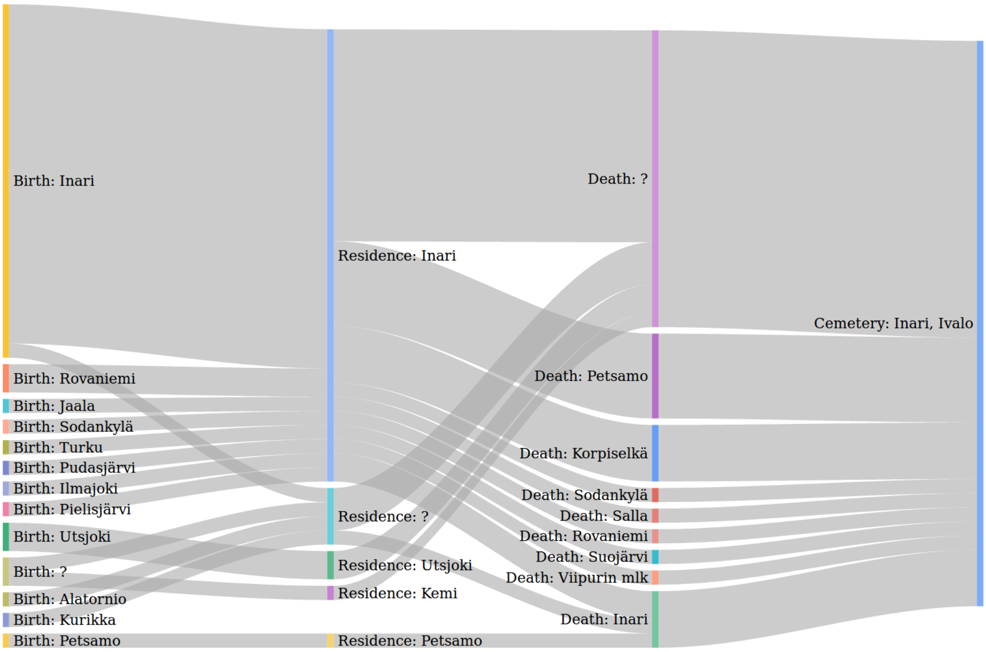


Figure 3.3. Life paths of 40 soldiers buried in the cemetery of Inari in Ivalo.

Publication VI presents restructured person home pages in the persons perspective, which reassembles soldier biographies by combining information contained in the person instances and in various person records. Differing values for every personal detail or activity are grouped together and shown on consecutive rows. Information sources are explicitly shown, whenever they are known, as the information can be contradictory in different sources within a prisoner record [108], or between different person records, or between person records and a person instance originating from other sources than the person records. The perspective serves citizens and researchers who are interested in finding information about a person's involvement in the war.

Publication VII describes the SPARQL Faceter tool, which provides the faceted search functionality for four of the WarSampo perspectives: Casualties, Photographs, War Cemeteries, and Prisoners of War. The tool is highly customizable, and can provide faceted search functionality over an arbitrary SPARQL endpoint. The data processing and handling of the facets happens on the client-side.

An asynchronous SPARQL query is sent from the user's web browser to the SPARQL endpoint each time a user makes a selection in the facets. The SPARQL endpoint returns the results of the query to the user's browser, which does additional processing of the data before displaying the new results to the user. The system works well even with the large casualty dataset, consisting of ca. 2.4 million triples, as pagination is used to limit the amount of results that are queried and displayed at a time.

Visualizing the Linked Data via SPARQL Endpoint. Publication VIII demonstrates the end-user use of the data directly from a SPARQL endpoint. For example, a user can query the daily casualties of a single military unit and all of its subunits. The results can be plotted with, e.g., the online YASGUI [174] tool, enabling easily visualizing the results. This way, a user can draw histograms with data directly obtained from the WarSampo SPARQL endpoint.

3.4 Maintaining Military Historical Linked Data

Research question 4 is concerned with maintaining the information contained in a Linked Data Cloud (LDC) in the domain of military history.

RQ4. How can the interlinked data and datasets be maintained?

The flexibility of the RDF data model provides various change propagation scenarios, where changes in one entity need to be taken into account in elsewhere due to links between entities, or the links would become invalidated. As ontologies are rarely static, this is a known problem in maintaining Linked Data.

The state of the art in Linked Data maintenance depends on the type of totality that is being focused on. Typically the distributed nature of the Semantic Web forces systems to react to external changes, which need to be first noticed, and then evaluated what has changed, and deciding whether the changes provoke a need to propagate the changes to the system in question. If the changes need to be propagated, then depending on the used approach, and the type of change, different actions are undertaken.

Publication IX addresses dataset maintenance on a LDC level, which improves on the state of the art in Linked Data maintenance by providing a practical scenario with a proposal for a solution in the case of a centrally managed LDC. The proposed solution is evaluated by demonstrating its use in maintaining the WarSampo knowledge graph. A LDC consists of a set of graphs, which can be differentiated into two major categories: DOs, and MDSs. DOs define the concepts used in populating the MDSs, and are shared by them. A set of DOs in an application domain is considered an *ontology infrastructure*. From a data management point of view, DOs, MDSs and mappings between graphs differ from each other.

The change propagation between graphs depends on whether the changed graph is a DO or an MDS, and whether a referencing graph is a DO or an MDS. The WarSampo ontology infrastructure is not static, but is maintained and extended to better represent the military history domain. As WarSampo heavily employs probabilistic entity linking to DOs, changes to a DO may invalidate existing entity linking, and the linking needs to be redone.

For example, maintenance of the War Cemeteries involves two scenarios of change propagation: 1) from DO to MDS, as the cemeteries are modeled as part of the places DO. If the cemetery data is updated, the linkage from the death records need to adjust to the change. 2) From MDS to DO, if the death records MDS, which references the cemeteries, changes, the cemeteries in the places DO may need to be adjusted.

As the prisoner data is maintained, new property values may be added. If new values are added to a property that is linked to a DO, the change should be propagated also to the DO, if a value is missing from it. This is the case, if for example a new occupation is added to the data, which is not present in the occupation DO. When a new person record is added to the register, the changes will propagate to the person DO, either through mapping, or through the creation of a new person instance. The person records are mapped to the person DO using probabilistic record linkage. The linking should be redone if the person DO changes to prevent broken or missing links.

A key lesson in iteratively building and maintaining the WarSampo dataset is that all data transformations and linking should be made into repeatable, automated processes, to be able to handle many change propagation scenarios automatically. The transformation processes should be built using a modular structure, to be maintainable and reusable. In a dynamic LDC, the entity linking processes need to be adaptable to changes.

A LDC that uses a complex data model, based on e.g., CIDOC CRM, will be difficult to maintain in RDF format. For complex DOs and MDSs, it is easier to update the data in simpler source formats, and maintain the data transformation processes that build the graphs. Simple independent DOs can be maintained directly in RDF format, whereas more complex DOs, e.g., people, require a different approach.

In Publication I, a data transformation pipeline is presented to solve the main change propagation scenarios in WarSampo. Processes in the pipeline take source datasets as input, transform data into RDF, and link entities to DOs. This automatically handles most of the change propagation scenarios, and easily prevents the linking between graphs from becoming inconsistent. The general idea is 1) first transforming the DOs, 2) then transforming datasets which both link to the person DO and create new person instances, and 3) then transforming and linking datasets that only make references to the DOs.

3.5 Results Summary

In this section, the research questions are revisited and summarized results presented.

1. How can wars be modeled and represented as data?

Military history is a promising use-case for Linked Data, facilitating the representation of heterogeneous, distributed, and conceptually interconnected information. Entities are given an identity and an identifier that can be shared between all the involved parties, and information is enriched just by making references to the identifiers of the shared entities.

As wars can be seen as sequences of events, event-based modeling is a natural framework for representing wars. CIDOC CRM is a widely used standard in the cultural heritage domain, providing an interoperable conceptual framework for event-based modeling.

In WarSampo, the CIDOC CRM has been extended to represent the military historical domain. Key extensions are subclasses of the CRM event class, that are used to present different events relating to the war, like battles, bombardments, political activity, and events of the actors participating in the war, like births, joinings to military units, troop movements, dissolutions, and promotions. Similarly the CRM properties are extended for the military historical domain. Detailed information sources for individual pieces of information are modeled as RDF Reifications. The data model can be easily extended as needed.

2. How can heterogeneous military historical data be harmonized and integrated from distributed sources?

Harmonization of the data requires interoperability on three distinct levels: 1) syntactic interoperability, 2) semantic interoperability, and 3) semantic disambiguation.

Using Linked Data provides the syntactic interoperability of heterogeneous datasets. By using CRM and a shared ontology infrastructure, the heterogeneous source datasets can be reconciled semantically to refer to shared entities instead of referring to entities by names. E.g., instead of referring to a person by their name, the person can be referred to by a URI, to make an unambiguous reference to a certain person.

A repeatable data transformation pipeline is used for building the majority of the knowledge graph 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. The semantic disambiguation is implemented using various NEL implementations to link resources to the DOs, and probabilistic record linkage to disambiguate people from different sources.

The entity linking enriches the metadatasets. For the entities in an MDS, the DO may contain further information about the linked concept, such as in the case of AMMO occupations, where linking a person to an occupation concept also enriches the person with information about his social status through the occupation. The DOs can also be used to provide contextual information through the entity linking, e.g., people with the same occupation.

3. How can military historical Linked Data be searched, browsed, analyzed, and visualized on web user interfaces?

WarSampo is targeted at all military historical data consumers: 1) academic researchers, 2) military history enthusiasts, and 3) private citizens. The WarSampo LOD service publishes all information as LOD, so researchers can download the data and process it further, or query and visualize it directly with SPARQL. The WarSampo portal provides user-friendly applications for all the user groups to search, browse, analyze, and visualize the data.

The portal consists of nine different application perspectives, that all provide user interfaces for searching and studying a certain part of the data. An important search paradigm employed in the user interfaces is faceted search, used in 5 perspectives. The perspectives show results mostly in tables, various visualizations, spatio-temporal views, and entity home pages. All of the key entities in the data have their own home pages within the perspectives. The perspectives are interlinked, by showing links to related entity home pages in other perspectives.

4. How can the interlinked data and datasets be maintained?

A Linked Data Cloud consists of a set of graphs, which can be differentiated into domain ontologies and metadatasets. From a data management point of view, DOs, MDSs, and mappings between graphs are different from each other, and changes in each produce different change propagation scenarios. The WarSampo ontology infrastructure is maintained and extended as needed to better represent the military history domain. As WarSampo uses mostly probabilistic entity linking to DOs, changes to a DO may invalidate existing entity linking, and the linking needs to be redone.

A key lesson is that data transformations and linking should be made into repeatable, automated processes, to be able to handle many change propagation scenarios automatically. The transformation processes should be built using a modular structure, to be maintainable and reusable. In a dynamic LDC, the entity linking processes need to be adaptable to changes.

A repeatable data transformation pipeline is used to solve change propagation scenarios in WarSampo. Processes in the pipeline take source datasets as input, transform data into RDF, and link entities to DOs. This automatically handles most of the change propagation scenarios, and easily prevents the graphs from going out of sync with each other.

4. Discussion

Traditional, comparative evaluation of the developed methods, tools and implementations in this thesis is difficult. These developed artifacts provide novel solutions for the research problems described in Chapter 1. Evaluating research is generally difficult in the Semantic Web research area [20], and a particular difficulty is that the usefulness and usability of the systems depend on multiple factors: the quality of heterogeneous source data used, the software used for data handling, and the user interfaces built for the data [211].

The following criteria have been used to evaluate the research of this thesis: 1) theoretical implications, 2) practical implications, 3) reliability, and 4) validity. In the following, the research of this thesis is evaluated against that criteria. Finally, recommendations for further research are presented.

4.1 Theoretical Implications

In comparison to earlier research on modeling and representing military history [42, 28, 216, 32, 59, 152, 45, 53, 2, 31, 210, 2, 49, 210], this thesis extends the widely used CIDOC CRM to the domain of military history as Linked Data. The existing scientific knowledge is advanced by presenting the WarSampo data model as a useful artifact, while demonstrating its applicability in practice in several case studies of populating different parts of the WarSampo knowledge graph, and using the data in the WarSampo portal. The created data model, and the populated knowledge graph are published as Linked Open Data, that can be used and further extended by anyone.

A repeatable data transformation pipeline is used for building the majority of the knowledge graph from the source datasets, aligning the data and linking entities in the process. This allows the domain experts to maintain the data in the original format and the changes can be integrated by recreating the whole knowledge graph. Time requirement of running the data transformation pipeline is a few hours, causing a minor delay in deploying data updates.

The WarSampo portal provides nine different perspectives on the data, combining multiple search paradigms, as opposed to many cultural heritage portals, which only provide a single view or search perspective of their contents [166, 189, 28, 31, 210]. The state of the art in the military history domain [152, 145, 28, 189, 210] is improved by providing event-based spatio-temporal user interfaces depicting the whole war as events. Additionally, other perspectives enable searching and browsing related resources like people, places, and photographs. The high level of interlinking within the knowledge graph is used to provide links between the different perspectives. In the WarSampo portal, the different perspectives are supported without modifying the data, but by only adjusting the data queries sent to the open SPARQL endpoint. WarSampo is the first large scale system for serving and publishing WW2 LOD on the Web.

Maintenance of Linked Data involves ontology evolution and Linked Data Dynamics, which have been studied in previous research in different contexts [131, 156, 224, 192, 131, 64, 164, 8, 101, 204, 136, 169]. This thesis improves the state of the art by providing the observed change propagation scenarios with a proposal for a practical solution in the case of a centrally managed Linked Data Cloud. A typology of change propagation is presented for describing the different scenarios. The WarSampo ontology infrastructure is not static, but is maintained and extended to better represent the military history domain.

4.2 Practical Implications

The WarSampo portal provides useful information for all citizens interested in the wars. The user interfaces provide different perspectives that a user can do searches on and browse different parts of the whole knowledge graph. The interlinking of resources makes the data richer than the individual datasets that a citizen could access through public memory institutions.

Military history enthusiasts and academic researchers can search and browse the data to focus on different parts according to their interests, like places or military units. They can also search for irregularities or patterns in the data through the user interfaces or by downloading the dataset and studying it with external tools. The implemented entity linking enables information retrieval based on the DOs, e.g., enabling a user to query everything in the data that has some relation to a specific place or a person.

Memory institutions, e.g., the National Archives of Finland, benefit from the results of this thesis by being able to host data in WarSampo, and in the future publish new datasets there, instead of building their own services. In the case of integrating the latest WarSampo dataset, the prisoners of war, WarSampo was chosen as the primary data publication platform by the stakeholders, which include the National Archives of Finland, and the Association for Cherishing the Memory of the Dead of the War. In addition, the lessons learned in WarSampo are valuable for an organization deciding to build their own system for historical information.

The WarSampo knowledge graph is published on the Linked Data Finland [88] platform, where it is openly available for use via a SPARQL endpoint, with the Creative Commons Attribution 4.0 license[1]. The WarSampo dataset page[2] contains human-readable information about the dataset and the SPARQL endpoint[3] serves all WarSampo data. A Fuseki[4] SPARQL Server is used for storing and serving the linked data. The used URIs are dereferenceable and provide information about the resources for both human and machine users, integrating the knowledge graph into the Semantic Web. By publishing openly shared ontologies and data about WW2 for everybody to use in annotations, hopefully future interoperability problems can be prevented.

The Casualties dataset in WarSampo has already been used as a basis for a popular Finnish WW2 portal, Sotapolku. Additionally, Wikidata has linked some Finnish person instances to WarSampo with a distinct WarSampo property, e.g., the commander-in-chief C. G. E. Mannerheim[5] is annotated with a WarSampo identifier.

Parts of the knowledge graph, especially the Places domain ontology and historical maps have been reused in the Finnish historical place and map service Hipla[6] as geo-gazetteers [96] and in the popular NameSampo service[7] for toponomastic research [97]. The AMMO occupation ontology has been re-used in the Finnish War Victims 1914–1922 project [172] and in a knowledge graph of historical Finnish academic people [124]. Finally, the knowledge graph was used for enriching data in the external semantic web applications *Norssi High School Alumni* [93] and *BiographySampo* [94].

4.3 Reliability and Validity

Realiability measures the consistency of the results and the consistency and stability of the research process over time and across researchers and

[1] `https://creativecommons.org/licenses/by/4.0/`

[2] `http://www.ldf.fi/dataset/warsampo/`

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

[4] `http://jena.apache.org/documentation/serving_data/`

[5] `https://www.wikidata.org/wiki/Q152306`

[6] `http://hipla.fi`

[7] `http://nimisampo.fi`

methods. The objectives of this study and the research questions are presented in Chapter 1, and the research questions are revisited in Chapter 3, when the results of the thesis are presented. The developed artifacts are presented in Chapter 3 and discussed in more details in the referenced publications, providing enough detail to support the repeatability of the research. The objectivity of the research is supported by the explicitly presented research methods, source datasets, and involved organizations. The author of the thesis has no competing interests or personal biases regarding the presented research that might have affected the research process.

Internal validity refers to the degree to which it is possible to draw conclusions from the observations. The developed artifacts meet the objectives set in Chapter 1, as is discussed in Chapter 3.

WarSampo is a part of the global LOD cloud[8] and was awarded with the LODLAM Challenge Open Data Prize in 2017[9]. The WarSampo knowledge graph has been accessed and used by more than 660,000 end users through the WarSampo portal, equivalent to more than 10% of the population of Finland. Over 400 end users have sent written feedback, mostly through the portal's feedback form. The feedback mostly concerns corrections to the data contents, usually of the details of a user's fallen relative. This suggests that most of the users are able to use the portal to find the information they are interested in, and are not unsatisfied with the experience of using the portal, as the comments usually do not take any stance on whether the portal is good or not.

The named entity linking [83] and person record linkage evaluation results presented in Chapter 3 are good enough to be useful in practice, as is demonstrated in the WarSampo portal.

External validity refers to the degree to which findings can be generalized beyond the setting in which they have been tested. The number of datasets integrated in the research demonstrate that the methods can be generalized to different contexts.

The scalability of the system in terms of concurrent users presents one limitation of the system. In 2017, a sudden peak in the public interest toward WarSampo made the service unavailable to users for some hours. The server capacity was increased to stabilize the situation and the system has since been quite stable. However, there are always limits to the concurrent users and this is related to much of the technical implementation of the server architecture like the hardware, the triple store used, caching, and other related software. During the 2017 peak, the system was hosted on a physical web server, which could not be scaled up to meet the demands. Currently, the system is hosted on a Kubernetes container orchestration system with docker containers, being able to automatically scale up as the

[8] http://linkeddata.org

[9] https://pro.europeana.eu/page/issue-7-lodlam

user demand increases [41].

As the data is gathered from more sources and the knowledge graph grows in size, scalability could become an issue also in the data transformation pipeline. The NEL processes and especially the person record linkage use plenty of computational resources. Currently the record linkage implementation in the pipeline is able to find links between the 4450 prisoner records and the 99,700 pre-existing person instances in the WarSampo actor ontology in a few hours on a modern desktop computer. The approach would probably need to be revised if the actor ontology would be considerably larger.

Even if the geographical scope of the research is Finland, there are no reason why the proposed methods and data model would not be directly applicable to other countries, provided that there would be data available to populate the crucial classes of the data model, like places, events, and people. Already, the integrated prisoners of war dataset contains data from Russian archives written in Russian language. The methods and data model could be applicable to another temporal scope, e.g., WW1.

4.4 Recommendations for Further Research

The research presented in this thesis provides a demonstration of how a deeper understanding about military history can be achieved through data integration, harmonization, and linking. The data contents however present only a small amount of all the actually relevant information that would be available in different archives and other data sources. The presented data model can be extended as needed to widen the scope of the data contents, without having to alter the existing parts of the metadata schema and ontology infrastructure. As more and more data sources are being digitized, there would be plenty of opportunities for studying various data integration cases in the future.

A topic for future research is combining information from different countries taking part in the war. This would enable painting a more global picture of events progressing geographically on a timeline, perhaps showing differences in the military history narratives of different countries.

In addition, the maintenance of data in RDF format remains a topic worth researching, as well as the possibility of harnessing the interest of the wider public by crowdsourcing contents that citizens have in their family belongings, etc. A collaborative platform for maintaining the data contents, like the one used in the Wikidata project [61], could provide fruitful for maintaining the data together by interested volunteers and professional historians. RDF validation is expected to become a fundamental enabler for data quality and interoperability [113]. Validating the whole knowledge graph and integrating validation to the WarSampo data transformation pipeline should be studied in the future. ShEx could be used to both validate the data and document the data model for data producers and consumers [114].

This thesis has not ventured deep into the modeling of tangible military historical cultural heritage although elements of this are captured in the cases of Heroes Cemeteries, prisoners of war camps, and medals. There would be more to explore in, e.g., military vessels, aircraft, weapons, and fortifications.

Bibliography

[1] AHLERS, D., AND BOLL, S. Location-based Web Search. In *The Geospatial Web*, A. Scharl and K. Tochtermann, Eds., Advanced Information and Knowledge Processing. Springer, London, 2009, pp. 55–66.

[2] ALEXIEV, V., NIKOLOVA, I., AND HATEVA, N. Semantic Archive Integration for Holocaust Research. The EHRI Research Infrastructure. *Umanistica Digitale 3*, 4 (2019).

[3] ALIAGA, D. G., BERTINO, E., AND VALTOLINA, S. DECHO - A Framework for the Digital Exploration of Cultural Heritage Objects. *Journal on Computing and Cultural Heritage (JOCCH) 3*, 3 (2011), 12.

[4] ALMA'AITAH, W. Z., TALIB, A. Z., AND OSMAN, M. A. Opportunities and challenges in enhancing access to metadata of cultural heritage collections: a survey. *Artificial Intelligence Review* (Oct 2019).

[5] ANDANY, J., LÉONARD, M., AND PALISSER, C. Management Of Schema Evolution In Databases. In *VLDB '91 Proceedings of the 17th International Conference on Very Large Data Bases* (September 1991), G. M. Lohman, A. Sernadas, and R. Camps, Eds., Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 161–170.

[6] AUDI, R. *Epistemology: A Contemporary Introduction to The Theory of Knowledge*. Routledge, 1998.

[7] AUER, S., BIZER, C., KOBILAROV, G., LEHMANN, J., CYGANIAK, R., AND IVES, Z. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*. Springer Berlin Heidelberg, 2007, pp. 722–735.

[8] AUER, S., DALAMAGAS, T., PARKINSON, H., BANCILHON, F., FLOURIS, G., SACHARIDIS, D., BUNEMAN, P., KOTZINOS, D., STAVRAKAS, Y., CHRISTOPHIDES, V., PAPASTEFANATOS, G., AND THIVEOS, K. Diachronic Linked Data: Towards Long-Term Preservation of Structured Interrelated Information. In *WOD'12: Proceedings of the First International Workshop on Open Data* (Nantes, France, May 2012), ACM, New York, NY, USA, pp. 31–39.

[9] AY, S. A., ZIMMERMANN, R., AND KIM, S. H. Viewable Scene Modeling for Geospatial Video Search. In *Proceedings of the 16th ACM international conference on Multimedia* (October 2008), ACM, New York, NY, USA, pp. 309–318.

[10] BAKER, T., BECHHOFER, S., ISAAC, A., MILES, A., SCHREIBER, G., AND SUMMERS, E. Key choices in the design of Simple Knowledge Organization System (SKOS). *Journal of Web Semantics 20* (2013), 35 – 49.

[11] BANERJEE, J., KIM, W., KIM, H.-J., AND KORTH, H. F. Semantics and Implementation of Schema Evolution in Object-oriented Databases. In *Proceedings of the 1987 ACM SIGMOD International Conference on Management of Data* (San Francisco, CA, USA, 1987), SIGMOD '87, ACM, New York, NY, USA, pp. 311–322.

[12] BATINI, C., LENZERINI, M., AND NAVATHE, S. B. A Comparative Analysis of Methodologies for Database Schema Integration. *ACM computing surveys (CSUR) 18*, 4 (1986), 323–364.

[13] BAZZANELLA, B., BORTOLI, S., AND BOUQUET, P. Can Persistent Identifiers Be Cool? *International Journal of Digital Curation 8*, 1 (2013), 14–28.

[14] BECKER, C. What is Historiography? *The American Historical Review 44*, 1 (1938), 20–28.

[15] BERNARD-DONALS, M. *An Introduction to Holocaust Studies*. Routledge, 2016.

[16] BERNERS-LEE, T. Linked Data - Design Issues, 2006. `http://www.w3.org/DesignIssues/LinkedData.html`, [Accessed 14.10.2019].

[17] BERNERS-LEE, T., CHEN, Y., CHILTON, L., CONNOLLY, D., DHANARAJ, R., HOLLENBACH, J., LERER, A., AND SHEETS, D. Tabulator: Exploring and Analyzing linked data on the Semantic Web. In *Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI06)* (2006), p. 159.

[18] BERNERS-LEE, T., HENDLER, J., AND LASSILA, O. The Semantic Web. *Scientific American 284*, 5 (2001), 28–37.

[19] BERNSTEIN, A., HENDLER, J., AND NOY, N. A New Look at the Semantic Web. *Communications of the ACM, New York, NY, USA 59*, 9 (Aug 2016), 35–37.

[20] BERNSTEIN, A., AND NOY, N. Is This Really Science? The Semantic Webber's Guide to Evaluating Research Contributions. Tech. rep., University of Zurich, Department of Informatics (IFI), 2014. Technical Report No. IFI-2014.02.

[21] BIDDLE, T. D., AND CITINO, R. M. The Role of Military History in the Contemporary Academy. *Foreign Policy Research Institute Footnotes* (February 2015), 1–6. `https://www.fpri.org/docs/society_for_mil_hist_whit_paper.pdf`, [Accessed 26.11.2019].

[22] BIKAKIS, N., AND SELLIS, T. Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art. In *Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference* (Bordeaux, France, March 2016), T. Palpanas and K. Stefanidis, Eds., vol. 1558, CEUR Workshop Proceedings, Aachen, Germany.

[23] BIZER, C., HEATH, T., AND BERNERS-LEE, T. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems 5*, 3 (2009), 1–22.

[24] BIZER, C., AND SCHULTZ, A. The R2R Framework: Publishing and Discovering Mappings on the Web. In *Proceedings of the First International Workshop on Consuming Linked Data (COLD2010)* (Shanghai, China, November 2010), O. Hartig, A. Harth, and J. Sequeda, Eds., vol. 665, CEUR Workshop Proceedings, Aachen, Germany.

[25] BLACK, J. *Rethinking Military History*. Routledge, 2004.

[26] BONTCHEVA, K., AND ROUT, D. Making Sense of Social Media Streams through Semantics: a Survey. *Semantic Web – Interoperability, Usability, Applicability 5*, 5 (2014), 373–403.

[27] BOONSTRA, O., BREURE, L., AND DOORN, P. Past, Present and Future of Historical Information Science. *Historical Social Research 29*, 2 (2004), 4–132.

[28] BOUKHELIFA, N., BRYANT, M., BULATOVIĆ, N., ČUKIĆ, I., FEKETE, J.-D., KNEŽEVIĆ, M., LEHMANN, J., STUART, D., AND THIEL, C. The CENDARI Infrastructure. *Journal on Computing and Cultural Heritage (JOCCH) 11*, 2 (2018), 8.

[29] BOYD, J. H., GUIVER, T., RANDALL, S. M., FERRANTE, A. M., SEMMENS, J. B., ANDERSON, P., AND DICKINSON, T. A Simple Sampling Method for Estimating the Accuracy of Large Scale Record Linkage Projects. *Methods of Information in Medicine 55*, 03 (2016), 276–283.

[30] BRAY, T., PAOLI, J., MALER, E., YERGEAU, F., AND COWAN, J., Eds. *Extensible Markup Language (XML) 1.1 (Second Edition)*. World Wide Web Consortium (W3C), 2006. https://www.w3.org/TR/2006/REC-xml11-20060816/, [Accessed 14.10.2019].

[31] BRAZZO, L., AND MAZZINI, S. Open Memory Project, April 2015. https://www.bygle.net/wp-content/uploads/2015/04/Open-Memory-Project_3-1.pdf, [Accessed 4.11.2019].

[32] BROWELL, G. From Linked Open Data to Linked Open Knowledge. In *Digital Information Strategies: From Applications and Content to Libraries and People*, D. Baker and W. Evans, Eds., Chandos Digital Information Review Series. Chandos Publishing, 2016.

[33] BUITINCK, L., AND MARX, M. Two-Stage Named-Entity Recognition Using Averaged Perceptrons. In *Natural Language Processing and Information Systems* (2012), G. Bouma, A. Ittoo, E. Métais, and H. Wortmann, Eds., Springer Berlin Heidelberg, pp. 171–176.

[34] BULST, N. Prosopography and the computer: Problems and possibilities. In *History and computing*, P. Denley, S. Fogelvik, and C. Harvey, Eds., vol. 2. Manchester University Press, 1989, pp. 12–18.

[35] BUNESCU, R. C., AND PASCA, M. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)* (Trento, Italy, April 2006), vol. 6, Association for Computational Linguistics, pp. 9–16.

[36] BURDICK, A., DRUCKER, J., LUNENFELD, P., PRESNER, T., AND SCHNAPP, J. *Digital Humanities*. The MIT Press, 2012.

[37] BURROWS, T., BRIX, A., EMERY, D., FRAAS, A. M., HYVÖNEN, E., IKKALA, E., KOHO, M., LEWIS, D., MYKING, S., RANSOM, L., THOMSON, E. C., TUOMINEN, J., WIJSMAN, H., AND WILCOX, P. Linked Open Data Vocabularies and Identifiers for Medieval Studies. In *Proceedings of Digital Humanities in Nordic Countries 5th Conference (DHN 2020)* (Riga, Latvia, October 2020). Accepted.

[38] BYRNE, K. Having Triplets – Holding Cultural Data as RDF. In *Proceedings of the ECDL 2008 Workshop on Information Access to Cultural Heritage* (Aarhus, Denmark, September 2008), University of Amsterdam, Information and Language Processing Systems group (ILPS).

[39] CAR, N. J., GOLODONIUC, P., AND KLUMP, J. The Challenge of Ensuring Persistency of Identifier Systems in the World of Ever-Changing Technology. *Data Science Journal 16* (2017), 13.

[40] CARR, D. *Time, Narrative, and History*. Indiana University Press, 1991.

[41] CASALICCHIO, E., AND PERCIBALLI, V. Auto-scaling of Containers: the Impact of Relative and Absolute Metrics. In *2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS* W)* (2017), IEEE, pp. 207–214.

[42] CENDARI PROJECT. CENDARI EDM Extension for WW1, Oct 2015. `https://repository.cendari.dariah.eu/es_AR/dataset/c42a9e3e-6615-41dc-be5e-d3bf74d37bde/resource/afe3ba32-de0a-4aaa-ab30-ce4f41eab159/download/cedmww1ontologyguidelines01.pdf`, [Accessed 20.10.2019].

[43] CHRISTEN, P. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[44] CITINO, R. M. Military Histories Old and New: A Reintroduction. *The American Historical Review 112*, 4 (2007), 1070–1090.

[45] COLLINS, T., MULHOLLAND, P., AND ZDRAHAL, Z. Semantic Browsing of Digital Collections. In *The Semantic Web – ISWC 2005: 4th International Semantic Web Conference* (Galway, Ireland, November 2005), Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, Eds., vol. 3729 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, pp. 127–141.

[46] COWLEY, R., AND PARKER, G. *The Reader's Companion to Military History*. Houghton Mifflin Harcourt (HMH), 1996.

[47] CYGANIAK, R., WOOD, D., LANTHALER, M., KLYNE, G., CARROLL, J. J., AND MCBRIDE, B. RDF 1.1 Concepts and Abstract Syntax. W3C recommendation, World Wide Web Consortium (W3C), 2014.

[48] DADZIE, A., AND ROWE, M. Approaches to Visualising Linked Data: A Survey. *Semantic Web – Interoperability, Usability, Applicability 2*, 2 (2011), 89–124.

[49] DAELEN, V. V. Data Sharing, Holocaust Documentation and the Digital Humanities: Introducing the European Holocaust Research Infrastructure (EHRI). *Umanistica Digitale 3*, 4 (2019).

[50] DAMIANO, R., AND LIETO, A. Ontological Representations of Narratives: a Case Study on Stories And Actions. In *2013 Workshop on Computational Models of Narrative* (2013), Schloss Dagstuhl Leibniz-Zentrum für Informatik, pp. 76–93.

[51] DAMLJANOVIC, D., AGATONOVIC, M., AND CUNNINGHAM, H. FREyA: An interactive way of querying Linked Data using natural language. In *Extended Semantic Web Conference* (2011), Springer, pp. 125–138.

[52] DAQUINO, M., MAMBELLI, F., PERONI, S., TOMASI, F., AND VITALI, F. Enhancing Semantic Expressivity in the Cultural Heritage Domain: Exposing the Zeri Photo Archive As Linked Open Data. *Journal on Computing and Cultural Heritage 10*, 4 (Jul 2017), 21:1–21:21.

[53] DE BOER, V., VAN DOORNIK, J., BUITINCK, L., MARX, M., AND VEKEN, T. Linking the Kingdom: Enriched Access To A Historiographical Text. In *Proceedings of the Seventh International Conference on Knowledge Capture (K-CAP 2013)* (Banff, Canada, June 2013), ACM, New York, NY, USA, pp. 17–24.

[54] DE LEEUW, D., BRYANT, M., FRANKL, M., NIKOLOVA, I., AND ALEXIEV, V. Digital Methods in Holocaust Studies: The European Holocaust Research Infrastructure. In *2018 IEEE 14th International Conference on e-Science (e-Science)* (2018), IEEE, pp. 58–66.

[55] DIMOU, A., SANDE, M. V., SLEPICKA, J., SZEKELY, P., MANNENS, E., KNOBLOCK, C., AND WALLE, R. V. D. Mapping Hierarchical Sources into RDF using the RML Mapping Language. *Proceedings - 2014 IEEE International Conference on Semantic Computing, ICSC 2014* (2014), 151–158.

[56] DOAN, A., AND HALEVY, A. Y. Semantic Integration Research in the Database Community: A Brief Survey. *AI Magazine - Special Issue on Semantic Integration 26*, 1 (2005), 83–83.

[57] DOERR, M. The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine 24*, 3 (2003), 75–75.

[58] DOMINOWSKA, A., HYTTINEN, E., IVANICS, P., KOHO, M., PIKKANEN, I., AND TURUNEN, R. Hiding in Plain Sight: Poetry in Newspapers and How to Approach it. *Human IT: Journal for Information Technology Studies as a Human Science 14*, 2 (2019), 145–171.

[59] EDELSTEIN, J., GALLA, L., LI-MADEO, C., MARDEN, J., RHONEMUS, A., AND WHYSEL, N. Linked Open Data for Cultural Heritage: Evolution of an Information Technology. Tech. rep., Columbia University Libraries, Libraries and Information Services, 2013. `https://academiccommons.columbia.edu/doi/10.7916/D8G44ZTM/download`, [Accessed 12.11.2019].

[60] ELO, K., AND KLEEMOLA, O. SA-kuva-arkistoa louhimassa: Digitaaliset tutkimusmenetelmät valokuvatutkimuksen tukena. In *Digitaalinen humanismi ja historiatieteet*, no. 12 in Historia mirabilis. Turun historiallinen yhdistys, 2016, pp. 151–190.

[61] ERXLEBEN, F., GÜNTHER, M., KRÖTZSCH, M., MENDEZ, J., AND VRANDEČIĆ, D. Introducing Wikidata to the Linked Data Web. In *The Semantic Web – ISWC 2014: 13th International Semantic Web Conference* (Riva del Garda, Italy, October 2014), P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, and C. Goble, Eds., vol. 8796 of *Lecture Notes in Computer Science*, Springer, Cham, pp. 50–65.

[62] FARRELL, J., AND NEZLEK, G. S. Rich Internet Applications: The Next Stage of Application Development. In *2007 29th International Conference on Information Technology Interfaces* (2007), IEEE, pp. 413–418.

[63] FLUIT, C., SABOU, M., AND VAN HARMELEN, F. Supporting User Tasks through Visualisation of Light-weight Ontologies. In *Handbook on Ontologies*. Springer, Berlin, Heidelberg, 2004, pp. 415–432.

[64] FROSTERUS, M., TUOMINEN, J., PESSALA, S., AND HYVÖNEN, E. Linked Open Ontology Cloud: Managing a System of Interlinked Cross-domain Light-weight Ontologies. *International Journal of Metadata, Semantics and Ontologies 10*, 3 (2015), 189–201.

[65] GAL, A., ANABY-TAVOR, A., TROMBETTA, A., AND MONTESI, D. A framework for modeling and evaluating automatic semantic reconciliation. *The VLDB Journal – The International Journal on Very Large Data Bases 14*, 1 (2005), 50–67.

[66] GAL, A., MODICA, G., JAMIL, H., AND EYAL, A. Automatic Ontology Matching Using Application Semantics. *AI magazine 26*, 1 (2005), 21–21.

[67] GASBARRA, L., KOHO, M., JOKIPII, I., RANTALA, H., AND HYVÖNEN, E. An Ontology of Finnish Historical Occupations. In *The Semantic Web: ESWC 2019 Satellite Events* (Portorož, Slovenia, June 2019), P. Hitzler, S. Kirrane, O. Hartig, V. de Boer, M.-E. Vidal, M. Maleshkova, S. Schlobach, K. Hammar, N. Lasierra, S. Stadtmüller, K. Hose, and R. Verborgh, Eds., vol. 11762 of *Lecture Notes in Computer Science*, Springer, Cham.

[68] GIUNCHIGLIA, F., AND ZAIHRAYEU, I. Lightweight Ontologies. In *Encyclopedia of Database Systems*, L. LIU and M. T. ÖZSU, Eds. Springer US, 2009, pp. 1613–1619.

[69] GOLSHAN, B., HALEVY, A., MIHAILA, G., AND TAN, W.-C. Data Integration: After the Teenage Years. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems* (2017), ACM, pp. 101–106.

[70] GRAFF, H. J. The Shock of the "'New' (Histories)": Social Science Histories and Historical Literacies. *Social Science History 25*, 4 (2001), 483–533.

[71] GRAHAM, S., MILLIGAN, I., AND WEINGART, S. *Exploring Big Historical Data: The Historian's Macroscope*. Imperial College Press, 2015.

[72] GREGOR, S., AND HEVNER, A. R. Positioning and Presenting Design Science Research for Maximum Impact. *MIS quarterly 37*, 2 (June 2013), 337–355.

[73] GRUBER, T. R. A Translation Approach to Portable Ontology Specifications. *Knowledge acquisition 5*, 2 (1993), 199–220.

[74] GU, L., BAXTER, R., VICKERS, D., AND RAINSFORD, C. Record linkage: Current practice and future directions. *CSIRO Mathematical and Information Sciences Technical Report 3* (2003), 83.

[75] HACHEY, B., RADFORD, W., NOTHMAN, J., HONNIBAL, M., AND CURRAN, J. R. Evaluating Entity Linking with Wikipedia. *Artificial Intelligence 194* (January 2013), 130–150.

[76] HARTIG, O. Provenance Information in the Web of Data. In *Proceedings of the WWW2009 Workshop on Linked Data on the Web* (Madrid, Spain, April 2009), C. Bizer, T. Heath, T. Berners-Lee, and K. Idehen, Eds., vol. 538, CEUR Workshop Proceedings, Aachen, Germany.

[77] HARVEY, F., KUHN, W., PUNDT, H., BISHR, Y., AND RIEDEMANN, C. Semantic interoperability: A central issue for sharing geographic information. *The annals of regional science 33*, 2 (1999), 213–232.

[78] HAST, A., CULLHED, P., AND VATS, E. TexT–Text extractor tool for handwritten document transcription and annotation. In *Digital Libraries and Multimedia Archives: 14th Italian Research Conference on Digital Libraries, IRCDL 2018, Udine, Italy, January 25-26, 2018, Proceedings* (2018), G. Serra and C. Tasso, Eds., vol. 806 of *Communications in Computer and Information Science*, Springer International Publishing, pp. 81–92.

[79] HEARST, M., ELLIOTT, A., ENGLISH, J., SINHA, R., SWEARINGEN, K., AND YEE, K.-P. Finding The Flow in Web Site Search. *Communications of the ACM 45*, 9 (2002), 42–49.

[80] HEATH, T., AND BIZER, C. *Linked Data: Evolving the Web into a Global Data Space*, 1st ed., vol. 1 of *Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool, 2011.

[81] HEFLIN, J., AND HENDLER, J. Semantic Interoperability on the Web. In *Proceedings of Extreme Markup Languages 2000* (2000), Graphic Communications Association, pp. 111–120.

[82] HEINO, E. Sotahistorian kuvaaminen ja rikastaminen linkitettynä datana. Master's thesis, University of Helsinki, Department of Computer Science, June 2017.

[83] HEINO, E., TAMPER, M., MÄKELÄ, E., LESKINEN, P., IKKALA, E., TUOMINEN, J., KOHO, M., AND HYVÖNEN, E. Named Entity Linking in a Complex Domain: Case Second World War History. In *Proceedings, Language, Technology and Knowledge (LDK 2017)* (Galway, Ireland, June 2017), Springer, Cham, pp. 120–133.

[84] HERT, M., REIF, G., AND GALL, H. C. A Comparison of RDB-to-RDF Mapping Languages. In *Proceedings of the 7th International Conference on Semantic Systems* (Graz, Austria, 2011), I-Semantics '11, ACM, New York, NY, USA, pp. 25–32.

[85] HEVNER, A. R., MARCH, S. T., PARK, J., AND RAM, S. Design Science in Information Systems Research. *MIS quarterly 28*, 1 (2004), 75–105.

[86] HILDEBRAND, M., VAN OSSENBRUGGEN, J., AND HARDMAN, L. /facet: A Browser for Heterogeneous Semantic Web Repositories. In *The Semantic Web – ISWC 2006: 5th International Semantic Web Conference* (Athens, GA, USA, November 2006), I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. M. Aroyo, Eds., vol. 4273 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 272–285.

[87] HYVÖNEN, E. Publishing and Using Cultural Heritage Linked Data on the Semantic Web. *Synthesis Lectures on the Semantic Web: Theory and Technology 2*, 1 (2012).

[88] HYVÖNEN, E., TUOMINEN, J., ALONEN, M., AND 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* (Anissaras, Crete, Greece, May 2014), V. Presutti, E. Blomqvist, R. Troncy, H. Sack, I. Papadakis, and A. Tordai, Eds., vol. 8798 of *Lecture Notes in Computer Science*, Springer, Cham, pp. 226–230.

[89] HYVÖNEN, E. Semantic Portals for Cultural Heritage. In *Handbook on Ontologies*, S. Staab and R. Studer, Eds., 2nd ed. Springer, Berlin, Heidelberg, April 2009.

[90] HYVÖNEN, E. Reconciling Metadata: 2 Data Reconciliation. In *Reassembling the Republic of Letters in the Digital Age*, H. Hotson and T. Wallnig, Eds. Göttingen University Press, 2019, ch. 3.2, pp. 223–235.

[91] HYVÖNEN, E., HEINO, E., LESKINEN, P., IKKALA, E., KOHO, M., TAMPER, M., TUOMINEN, J., AND MÄKELÄ, E. Publishing Second World War History as Linked Data Events on the Semantic Web. In *Proceedings of Digital Humanities 2016, short papers* (Kraków, Poland, July 2016), pp. 571–573.

[92] HYVÖNEN, E., IKKALA, E., TUOMINEN, J., KOHO, M., BURROWS, T., RANSOM, L., AND WIJSMAN, H. A Linked Open Data Service and Portal for Pre-modern Manuscript Research. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN 2019)* (Copenhagen, Denmark, March 2019), C. Navarretta, M. Agirrezabal, and B. Maegaard, Eds., vol. 2364, CEUR Workshop Proceedings, Aachen, Germany, pp. 220–229.

[93] HYVÖNEN, E., LESKINEN, P., HEINO, E., TUOMINEN, J., AND SIROLA, L. 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)* (June 2017), Springer, Cham, pp. 113–119.

[94] HYVÖNEN, E., LESKINEN, P., TAMPER, M., RANTALA, H., IKKALA, E., TUOMINEN, J., AND KERAVUORI, K. BiographySampo - Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research. In *The Semantic Web: ESWC 2019 Satellite Events* (Portorož, Slovenia, June 2019), P. Hitzler, S. Kirrane, O. Hartig, V. de Boer, M.-E. Vidal, M. Maleshkova, S. Schlobach, K. Hammar, N. Lasierra, S. Stadtmüller, K. Hose, and R. Verborgh, Eds., vol. 11762 of *Lecture Notes in Computer Science*, Springer, Cham.

[95] HYVÖNEN, E., MÄKELÄ, E., KAUPPINEN, T., ALM, O., KURKI, J., RUOTSALO, T., SEPPÄLÄ, K., TAKALA, J., PUPUTTI, K., KUITTINEN, H., VILJANEN, K., TUOMINEN, J., PALONEN, T., FROSTERUS, M., SINKKILÄ, R., PAAKKARINEN, P., LAITIO, J., AND NYBERG, K. CultureSampo – A National Publication System of Cultural Heritage on the Semantic Web 2.0. In *The Semantic Web: Research and Applications, 6th European Semantic Web Conference, ESWC 2009* (Heraklion, Crete, Greece, May 31 - June 4 2009), L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, Eds., vol. 5554 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg.

[96] IKKALA, E., HYVÖNEN, E., AND TUOMINEN, J. An Ontology of World War II Places for Linking and Enriching Heterogeneous Historical Data Sources. In *17th International Conference of Historical Geographers (ICHG 2018), Book of Abstracts* (Warsaw, Poland, July 2018), no. 194.

[97] IKKALA, E., TUOMINEN, J., RAUNAMAA, J., AALTO, T., AINIALA, T., UUSITALO, H., AND HYVÖNEN, E. NameSampo: A Linked Open Data Infrastructure and Workbench for Toponomastic Research. In *GeoHumanities'18: Proceedings of the 2nd ACM SIGSPATIAL Workshop on Geospatial Humanities* (Seattle, WA, USA, November 2018), P. Murrieta and B. Martins, Eds., ACM, New York, NY, USA, pp. 2:1–2:9.

[98] IOANNOU, E., NIEDERÉE, C., AND NEJDL, W. Probabilistic Entity Linkage for Heterogeneous Information Spaces. In *International Conference on Advanced Information Systems Engineering* (2008), Springer, pp. 556–570.

[99] ISAAC, A., Ed. *Europeana Data Model Primer*. Europeana, 2013. https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf, [Accessed 3.11.2019].

[100] ISAAC, A., AND HASLHOFER, B. Europeana Linked Open Data – data.europeana.eu. *Semantic Web – Interoperability, Usability, Applicability 4*, 3 (2013), 291–297.

[101] KÄFER, T., ABDELRAHMAN, A., UMBRICH, J., O'BYRNE, P., AND HOGAN, A. Observing linked data dynamics. In *The Semantic Web: Semantics and Big Data* (2013), P. Cimiano, O. Corcho, V. Presutti, L. Hollink, and S. Rudolph, Eds., Springer Berlin Heidelberg, pp. 213–227.

[102] KATIFORI, A., HALATSIS, C., LEPOURAS, G., VASSILAKIS, C., AND GIANNOPOULOU, E. Ontology Visualization Methods – A Survey. *ACM Computing Surveys (CSUR) 39*, 4 (2007), 10.

[103] KINNUNEN, T., AND KIVIMÄKI, V., Eds. *Finland in World War II: history, memory, interpretations*. Brill, 2011.

[104] KIVIMÄKI, V. *Murtuneet mielet: taistelu suomalaissotilaiden hermoista 1939-1945*. WSOY, 2013.

[105] KIVIMÄKI, V., AND TEPORA, T. Meaningless Death or Regenerating Sacrifice? Violence and Social Cohesion in Wartime Finland. In *Finland in World War II : History, Memory, Interpretations*, T. Kinnunen and V. Kivimäki, Eds., vol. 69 of *History of Warfare*. Brill, 2012, pp. 233–275.

[106] KNOBLOCK, C. A., SZEKELY, P., AMBITE, J. L., GOEL, A., GUPTA, S., LERMAN, K., MUSLEA, M., TAHERIYAN, M., AND MALLICK, P. Semi-Automatically Mapping Structured Sources into the Semantic Web. In *Extended Semantic Web Conference* (2012), E. Simperl, P. Cimiano, A. Polleres, O. Corcho, and V. Presutti, Eds., Springer Berlin Heidelberg, pp. 375–390.

[107] KNOBLOCK, C. A., SZEKELY, P., FINK, E., DEGLER, D., NEWBURY, D., SANDERSON, R., BLANCH, K., SNYDER, S., CHHEDA, N., JAIN, N., RAJU KRISHNA, R., BEGUR SREEKANTH, N., AND YAO, Y. Lessons Learned in Building Linked Data for the American Art Collaborative. In *The Semantic Web – ISWC 2017: 16th International Semantic Web Conference* (Vienna, Austria, October 2017), C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, and J. Heflin, Eds., vol. 10588 of *Lecture Notes in Computer Science*, Springer, Cham, pp. 263–279.

[108] KOHO, M., HEINO, E., IKKALA, E., HYVÖNEN, E., NIKKILÄ, R., MOILANEN, T., MIETTINEN, K., AND SUOMINEN, P. Integrating Prisoners of War Dataset into the WarSampo Linked Data Infrastructure. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018)* (Helsinki, Finland, March 2018), E. Mäkelä, M. Tolonen, and J. Tuominen, Eds., vol. 2084, CEUR Workshop Proceedings, Aachen, Germany.

[109] KOHO, M., HEINO, E., LESKINEN, P., IKKALA, E., TAMPER, M., APAJALAHTI, K., TUOMINEN, J., MÄKELÄ, E., AND HYVÖNEN, E. WarSampo Knowledge Graph [Data set], Oct. 2019. `https://doi.org/10.5281/zenodo.3431121`, [Accessed 7.11.2019].

[110] KOHO, M., HEINO, E., OKSANEN, A., AND HYVÖNEN, E. Toffee - Semantic Media Search Using Topic Modeling and Relevance Feedback. In *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks* (Monterey, CA, USA, October 2018), M. van Erp, M. Atre, V. Lopez, K. Srinivas, and C. Fortuna, Eds., vol. 2180, CEUR Workshop Proceedings, Aachen, Germany.

[111] KOSKINEN-KOIVISTO, E., AND THOMAS, S. Lapland's Dark Heritage: Responses to the Legacy of World War II. In *Heritage in Action*, H. Silverman, E. Waterton, and S. Watson, Eds. Springer Cham, 2017, pp. 121–133.

[112] KOURIJOKI, A. Linkitetyn datan validointi ja korjaus. Master's thesis, Aalto University, Department of Computer Science, 2020. Accepted.

[113] LABRA GAYO, J. E., PRUD'HOMMEAUX, E., BONEVA, I., AND KONTOKOSTAS, D. *Validating RDF data*, vol. 16 of *Synthesis Lectures on The Semantic Web: Theory and Technology*. Morgan & Claypool Publishers, 2017.

[114] LABRA GAYO, J. E., PRUD'HOMMEAUX, E., SOLBRIG, H., AND RODRÍGUEZ, J. M. A. Validating and describing linked data portals using rdf shape expressions. In *Proceedings of the 1st Workshop on Linked Data Quality (LDQ 2014)* (Leipzig, Germany, September 2014), M. Knuth, D. Kontokostas, and H. Sack, Eds., no. 1215 in CEUR Workshop Proceedings, Aachen, Germany.

[115] LAKSHMANAN, L. V. S., AND SADRI, F. Interoperability on XML Data. In *The Semantic Web - ISWC 2003: Second International Semantic Web Conference* (Sanibel Island, FL, USA, October 2003), D. Fensel, K. Sycara, and J. Mylopoulos, Eds., vol. 2870 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, pp. 146–163.

[116] LAMBERT, P. S., ZIJDEMAN, R. L., VAN LEEUWEN, M. H. D., MAAS, I., AND PRANDY, K. The Construction of HISCAM: A Stratification Scale Based on Social Interactions for Historical Comparative Research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History 46*, 2 (2013), 77–89.

[117] LAUSEN, H., DING, Y., STOLLBERG, M., FENSEL, D., LARA HERNÁNDEZ, R., AND HAN, S.-K. Semantic web portals: state-of-the-art survey. *Journal of Knowledge Management 9*, 5 (2005), 40–49.

[118] LEFRANÇOIS, M., ZIMMERMANN, A., AND BAKERALLY, N. A SPARQL Extension for Generating RDF from Heterogeneous Formats. In *The Semantic Web* (2017), E. Blomqvist, D. Maynard, A. Gangemi, R. Hoekstra, P. Hitzler, and O. Hartig, Eds., Springer International Publishing, pp. 35–50.

[119] LEHMANN, J., ISELE, R., JAKOB, M., JENTZSCH, A., KONTOKOSTAS, D., MENDES, P. N., HELLMANN, S., MORSEY, M., VAN KLEEF, P., AUER, S., AND BIZER, C. DBpedia–A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web – Interoperability, Usability, Applicability 6*, 2 (2015), 167–195.

[120] LEINONEN, R.-M. Finnish narratives of the horse in World War II. *Animals and war: studies of Europe and North America* (2013), 123–150.

[121] LENTILÄ, R. Sodissa menehtyneiden tiedosto. In *Yhdessä Kestämme – Suomen Sotaveteraaniliitto ry 40 vuotta 29.9.1997*, S. Kärävä, A. Hartikka, E. Kosunen, J. Valve, A. Henttonen, and J. Ketola, Eds. Suomen Sotaveteraaniliitto ry, 1997, pp. 87–96.

[122] LESKINEN, J., AND JUUTILAINEN, A., Eds. *Jatkosodan pikkujättiläinen*. WSOY, Finland, 2005.

[123] LESKINEN, J., AND JUUTILAINEN, A., Eds. *Talvisodan pikkujättiläinen*, 4th ed. WSOY, Finland, 2006.

[124] LESKINEN, P., AND HYVÖNEN, E. Linked Open Data Service about Historical Finnish Academic People in 1640–1899. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020)* (2019), CEUR Workshop Proceedings, Aachen, Germany. Submitted.

[125] LESKINEN, P., MIYAKITA, G., KOHO, M., AND HYVÖNEN, E. Combining Faceted Search with Data-analytic Visualizations on Top of a SPARQL Endpoint. In *Proceedings of the Fourth International Workshop on Visualization and Interaction for Ontologies and Linked Data (VOILA 2018)* (Monterey, CA, USA, October 2018), V. Ivanova, P. Lambrix, S. Lohmann, and C. Pesquita, Eds., vol. 2187, CEUR Workshop Proceedings, Aachen, Germany.

[126] LINDQUIST, T., HYVÖNEN, E., TÖRNROOS, J., AND MÄKELÄ, E. Leveraging linked data to enhance subject access - A case study of the University of Colorado Boulder's World War I collection online. In *World Library and Information Congress: 78th IFLA General Conference and Assembly, Helsinki* (August 2012), IFLA.

[127] LITTLE, D. Philosophy of History. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., summer 2017 ed. Metaphysics Research Lab, Stanford University, 2017. `https://plato.stanford.edu/archives/sum2017/entries/history/`, [Accessed 7.11.2019].

[128] LITZ, B., LÖHDEN, A., HANNEMANN, J., AND SVENSSON, L. AgRelOn – An Agent Relationship Ontology. In *Metadata and Semantics Research* (2012), J. M. Dodero, M. Palomo-Duarte, and P. Karampiperis, Eds., Springer Berlin Heidelberg, pp. 202–213.

[129] LOPEZ, V., FERNÁNDEZ, M., STIELER, N., AND MOTTA, E. PowerAqua: Supporting Users in Querying and Exploring the Semantic Web Content. *Semantic Web – Interoperability, Usability, Applicability 3*, 3 (2012), 249–265.

[130] MAALI, F., CYGANIAK, R., AND PERISTERAS, V. A Publishing Pipeline for Linked Government Data. In *The Semantic Web: Research and Applications* (2012), E. Simperl, P. Cimiano, A. Polleres, O. Corcho, and V. Presutti, Eds., Springer Berlin Heidelberg, pp. 778–792.

[131] MAEDCHE, A., MOTIK, B., STOJANOVIC, L., STUDER, R., AND VOLZ, R. An Infrastructure for Searching, Reusing and Evolving Distributed Ontologies. In *Proceedings of the 12th international conference on World Wide Web* (Budapest, Hungary, 2003), WWW '03, ACM, New York, NY, USA, pp. 439–448.

[132] MANDEMAKERS, K., MOURITS, R. J., MUURLING, S., BOTER, C., VAN DIJK, I. K., MAAS, I., DE PUTTE, B. V., ZIJDEMAN, R. L., LAMBERT, P., VAN LEEUWEN, M. H., VAN POPPEL, F., AND MILES, A. *HSN standardized, HISCO-coded and classified occupational titles, release 2018.01.* IISG, Amsterdam, The Netherlands, 2018.

[133] MANOUSIS, P., VASSILIADIS, P., ZARRAS, A., AND PAPASTEFANATOS, G. Schema evolution for databases and data warehouses. In *European Business Intelligence Summer School* (2015), Springer, pp. 1–31.

[134] MARCH, S. T., AND SMITH, G. F. Design And Natural Science Research on Information Technology. *Decision support systems 15*, 4 (1995), 251–266.

[135] MARKEY, K., ATHERTON, P., AND NEWTON, C. An analysis of controlled vocabulary and free text search statements in online searches. *Online review 4*, 3 (1980), 225–236.

[136] MEIMARIS, M., PAPASTEFANATOS, G., PATERITSAS, C., GALANI, T., AND STAVRAKAS, Y. Towards a Framework for Managing Evolving Information Resources on the Data Web. In *Proceedings of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data (PROFILES 2014)* (Anissaras, Crete, Greece, March 2014), E. Demidova, S. Dietze, J. Szymanski, and J. Breslin, Eds., vol. 1151, CEUR Workshop Proceedings, Aachen, Germany.

[137] MEROÑO-PEÑUELA, A., ASHKPOUR, A., VAN ERP, M., MANDEMAKERS, K., BREURE, L., SCHARNHORST, A., SCHLOBACH, S., AND VAN HARMELEN, F. Semantic Technologies For Historical Research: A Survey. *Semantic Web – Interoperability, Usability, Applicability 6*, 6 (2015), 539–564.

[138] METILLI, D., BARTALESI, V., AND MEGHINI, C. A Wikidata-based tool for building and visualising narratives. *International Journal on Digital Libraries* (Jan 2019).

[139] MICHELFEIT, J., KNAP, T., AND NEČASKÝ, M. Linked Data Integration with Conflicts. *ArXiv abs/1410.7990* (2014).

[140] Miles, A., Matthews, B., Wilson, M., and Brickley, D. SKOS Core: Simple knowledge organisation for the Web. In *International Conference on Dublin Core and Metadata Applications* (2005), pp. 3–10.

[141] Mongiovì, M., Recupero, D. R., Gangemi, A., Presutti, V., Nuzzolese, A. G., and Consoli, S. Semantic Reconciliation of Knowledge Extracted from Text Through a Novel Machine Reader. In *Proceedings of the 8th International Conference on Knowledge Capture (K-CAP 2015)* (Palisades, NY, USA, 2015), K-CAP 2015, ACM, New York, NY, USA, pp. 25:1–25:4.

[142] Morillo, S. *What is Military History?* John Wiley & Sons, 2017.

[143] Moro, A., Raganato, A., and Navigli, R. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics 2* (2014), 231–244.

[144] Mulholland, P., Collins, T., and Zdrahal, Z. Story Fountain: Intelligent Support For Story Research And Exploration. In *Proceedings of the 9th international conference on Intelligent user interfaces* (2004), ACM, pp. 62–69.

[145] Mulholland, P., Collins, T., and Zdrahal, Z. Bletchley Park text: Using mobile and semantic web technologies to support the post-visit use of online museum resources. *Journal of Interactive Media in Education 24*, Specia (2005).

[146] Mustajoki, H. Kohtalo omissa käsissä: Suomen sodissa 1939–1945 itsensä surmanneiden sotilaiden omaisten asema vuosina 1939–1960. Master's thesis, University Of Helsinki, March 2010.

[147] Musto, C., Lops, P., Basile, P., de Gemmis, M., and Semeraro, G. Semantics-aware Graph-based Recommender Systems Exploiting Linked Open Data. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (Halifax, Nova Scotia, Canada, 2016), UMAP '16, ACM, New York, NY, USA, pp. 229–237.

[148] Mäkelä, E. *View-Based User Interfaces for the Semantic Web*. PhD thesis, Aalto University, School of Science and Technology, November 2010.

[149] Mäkelä, E. Combining a REST Lexical Analysis Web Service with SPARQL for Mashup Semantic Annotation from Text. In *The Semantic Web: ESWC 2014 Satellite Events* (Anissaras, Crete, Greece, May 2014), V. Presutti, E. Blomqvist, R. Troncy, H. Sack, I. Papadakis, and A. Tordai, Eds., vol. 8798 of *Lecture Notes in Computer Science*, Springer, Cham, pp. 226–230.

[150] Mäkelä, E., Hyvönen, E., and Ruotsalo, T. How to deal with massively heterogeneous cultural heritage data – lessons learned in CultureSampo. *Semantic Web – Interoperability, Usability, Applicability 3*, 1 (January 2012).

[151] Mäkelä, E., Lindquist, T., and Hyvönen, E. CORE - A Contextual Reader based on Linked Data. In *Proceedings of Digital Humanities 2016, Long Papers* (Kraków, Poland, July 2016), pp. 267–269.

[152] Mäkelä, E., Törnroos, J., Lindquist, T., and Hyvönen, E. WW1LOD: An application of CIDOC-CRM to World War 1 linked data. *International Journal on Digital Libraries 18*, 4 (nov 2017), 333–343.

[153] Nagypál, G., Deswarte, R., and Oosthoek, J. Applying the Semantic Web: The VICODI Experience in Creating Visual Contextualization for History. *Literary and Linguistic Computing 20*, 3 (2005), 327–349.

[154] NAVIGLI, R. Word Sense Disambiguation: A Survey. *ACM Computing Surveys (CSUR) 41*, 2 (2009), 10.

[155] NGUYEN, V., BODENREIDER, O., AND SHETH, A. Don't Like RDF Reification? Making Statements about Statements Using Singleton Property. In *WWW '14: Proceedings of the 23rd International Conference on World Wide Web* (Seoul, Korea, April 2014), Association for Computing Machinery, New York, NY, USA, pp. 759–770.

[156] NOY, N. F., AND KLEIN, M. Ontology Evolution: Not the Same as Schema Evolution. *Knowledge and information systems 6*, 4 (2004), 428–440.

[157] OREN, E., DELBRU, R., AND DECKER, S. Extending Faceted Navigation for RDF data. In *The Semantic Web – ISWC 2006: 5th International Semantic Web Conference* (Athens, GA, USA, November 2006), I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. M. Aroyo, Eds., vol. 4273 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 559–572.

[158] ÖZACAR, T., ÖZTÜRK, Ö., SALLOUTAH, L., YÜKSEL, F., ABDÜLBAKI, B., AND BILICI, E. A Semantic Web Case Study: Representing the Ephesus Museum Collection Using Erlangen CRM Ontology. In *Research Conference on Metadata and Semantics Research* (2017), Springer, pp. 202–210.

[159] PAN, J. Z. Resource Description Framework. In *Handbook on Ontologies*. Springer, Berlin, Heidelberg, 2009, pp. 71–90.

[160] PARENT, C., AND SPACCAPIETRA, S. Database Integration: The Key to Data Interoperability. *Advances in Object-Oriented Data Modelling* (2000).

[161] PEFFERS, K., TUUNANEN, T., ROTHENBERGER, M., AND CHATTERJEE, S. A Design Science Research Methodology for Information Systems Research. *J. Manage. Inf. Syst. 24*, January (2008), 45–77.

[162] PELLISSIER TANON, T., VRANDEČIĆ, D., SCHAFFERT, S., STEINER, T., AND PINTSCHER, L. From Freebase to Wikidata: The Great Migration. In *Proceedings of the 25th international conference on world wide web* (Montréal, Québec, Canada, 2016), WWW '16, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 1419–1428.

[163] PERONI, S., TOMASI, F., AND VITALI, F. Reflecting on the Europeana Data Model. In *Digital Libraries and Archives* (2013), M. Agosti, F. Esposito, S. Ferilli, and N. Ferro, Eds., Springer Berlin Heidelberg, pp. 228–240.

[164] PESSALA, S., SEPPÄLÄ, K., SUOMINEN, O., FROSTERUS, M., TUOMINEN, J., AND HYVÖNEN, E. MUTU: An Analysis Tool for Maintaining a System of Hierarchically Linked Ontologies. In *Proceedings of the ISWC 2011 Workshop Ontologies Come of Age in the Semantic Web (OCAS-2011)* (Bonn, Germany, October 2011), A. G. Castro, K. Baclawski, J. Bateman, C. Lange, and K. Viljanen, Eds., vol. 809, CEUR Workshop Proceedings, Aachen, Germany.

[165] PETERSON, D., GAO, S. S., MALHOTRA, A., SPERBERG-MCQUEEN, C. M., AND THOMPSON, H. S., Eds. *W3C XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes*. World Wide Web Consortium (W3C), 2012. `http://www.w3.org/TR/xmlschema11-2/`, [Accessed 14.10.2019].

[166] PETRAS, V., HILL, T., STILLER, J., AND GÄDE, M. Europeana – a Search Engine for Digitised Cultural Heritage Material. *Datenbank-Spektrum 17*, 1 (Mar 2017), 41–46.

[167] PIEDRA, N., TOVAR, E., COLOMO-PALACIOS, R., LOPEZ-VARGAS, J., AND ALEXANDRA CHICAIZA, J. Consuming and producing linked open data: the case of Opencourseware. *Program: electronic library and information systems 48*, 1 (2014), 16–40.

[168] POLLITT, A. S. The key role of classification and indexing in view-based searching. Tech. rep., Centre for Database Access Research,University of Huddersfield, 1998. http://www.ifla.org/IV/ifla63/63polst.pdf, [Accessed 8.11.2019].

[169] POPITSCH, N. P., AND HASLHOFER, B. DSNotify: Handling Broken Links in the Web of Data. In *Proceedings of the 19th international conference on World wide web* (Raleigh, NC, USA, 2010), WWW '10, ACM, New York, NY, USA, pp. 761–770.

[170] RAIMOND, Y., ABDALLAH, S. A., SANDLER, M. B., AND GIASSON, F. The Music Ontology. In *ISMIR 2007: Proceedings of the 8th International Conference on Music Information Retrieval* (2007), Austrian Computer Society.

[171] RANTALA, H., IKKALA, E., JOKIPII, I., KOHO, M., TUOMINEN, J., AND HYVÖNEN, E. WarVictimSampo 1914–1922: A Semantic Portal and Linked Data Service for Digital Humanities Research on War History. In *The Semantic Web: ESWC 2020 Satellite Events* (May 31 - June 4 2020). Accepted.

[172] RANTALA, H., JOKIPII, I., KOHO, M., IKKALA, E., TUOMINEN, J., AND HYVÖNEN, E. Building a Linked Open Data Portal of War Victims in Finland 1914-1922. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020)* (Riga, Latvia, October 2020). Accepted.

[173] RASILA, V. *Kansalaissodan sosiaalinen tausta*. Tammi, Helsinki, Finland, 1968.

[174] RIETVELD, L., AND HOEKSTRA, R. The YASGUI Family of SPARQL clients. *Semantic Web – Interoperability, Usability, Applicability 8*, 3 (2017), 373–383.

[175] ROSSI, G., SÁNCHEZ-FIGUEROA, F., AND FRATERNALI, P. Rich Internet Applications. *IEEE Internet Computing 14*, 03 (may 2010), 9–12.

[176] ROVERA, M. A Knowledge-Based Framework for Events Representation and Reuse from Historical Archives. In *The Semantic Web. Latest Advances and New Domains* (2016), H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S. P. Ponzetto, and C. Lange, Eds., Springer International Publishing, pp. 845–852.

[177] RUGGLES, D. F., AND SILVERMAN, H. From Tangible to Intangible Heritage. In *Intangible Heritage Embodied*, H. Silverman and D. F. Ruggles, Eds. Springer New York, New York, NY, USA, 2009, pp. 1–14.

[178] SAARELA, J., AND FINNÄS, F. Long-term mortality of war cohorts: The case of Finland. *European Journal of Population / Revue européenne de Démographie 28*, 1 (2012), 1–15.

[179] SACCO, G. M. Dynamic taxonomies: guided interactive diagnostic assistance. In *Encyclopedia of Healthcare Information Systems*, N. Wickramasinghe, Ed. Idea Group, 2007.

[180] SAHOO, S. S., HALB, W., HELLMANN, S., IDEHEN, K., THIBODEAU JR, T., AUER, S., SEQUEDA, J., AND EZZAT, A. A survey of current approaches for mapping of relational databases to RDF. *W3C RDB2RDF Incubator Group Report 1* (2009), 113–130.

[181] SANDERSON, R., AND VAN DE SOMPEL, H. Cool URIs and Dynamic Data. *IEEE Internet Computing 16*, 4 (2012), 76–79.

[182] SCHERP, A., FRANZ, T., SAATHOFF, C., AND STAAB, S. 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)* (Redondo Beach, CA, USA, September 2009), ACM, New York, NY, USA, pp. 137–144.

[183] SEITSONEN, O., AND HERVA, V.-P. "War junk" and Cultural Heritage: Viewpoints on the Second World War German Material Culture in the Finnish Lapland. In *War & Peace: Conflict and Resolution in Archaeology. Proceedings of the 45th Annual Chacmool Archaeology Conference.* (2017), A. K. Benfer, Ed., Chacmool Archaeology Association, University of Calgary.

[184] SHADBOLT, N., BERNERS-LEE, T., AND HALL, W. The Semantic Web Revisited. *IEEE intelligent systems 21*, 3 (2006), 96–101.

[185] SHADBOLT, N., O'HARA, K., BERNERS-LEE, T., GIBBINS, N., GLASER, H., HALL, W., AND M.C. SCHRAEFEL. Linked Open Government Data: Lessons from Data.gov.uk. *IEEE Intelligent Systems 27*, 3 (2012), 16–24.

[186] SHAW, R., TRONCY, R., AND HARDMAN, L. LODE: Linking Open Descriptions of Events. In *The Semantic Web. Fourth Asian Conference, ASWC 2009.* (2009), A. Gómez-Pérez, Y. Yu, and Y. Ding, Eds., vol. 5926 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, pp. 153–167.

[187] SHEN, W., WANG, J., AND HAN, J. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering 27*, 2 (2014), 443–460.

[188] SHNEIDERMAN, B., BYRD, D., AND CROFT, W. B. Clarifying search: A user-interface framework for text searches. *D-lib magazine 3*, 1 (1997), 18–20.

[189] SPRUGNOLI, R., MORETTI, G., AND TONELLI, S. LOD Navigator: Tracing Movements of Italian Shoah Victims. *Umanistica Digitale 3*, 4 (2019).

[190] STAAB, S., ANGELE, J., DECKER, S., ERDMANN, M., HOTHO, A., MAEDCHE, A., SCHNURR, H.-P., STUDER, R., AND SURE, Y. Semantic Community Web Portals. *Computer Networks 33*, 1-6 (2000), 473–491.

[191] STATISTICS FINLAND. *Classification of Occupations 1980*. Käsikirjoja / Tilastokeskus. Statistics Finland, Helsinki, Finland, 1981.

[192] STOJANOVIC, L., MAEDCHE, A., MOTIK, B., AND STOJANOVIC, N. User-Driven Ontology Evolution Management. In *International Conference on Knowledge Engineering and Knowledge Management* (2002), Springer, pp. 285–300.

[193] STOJANOVIC, N., MAEDCHE, A., STAAB, S., STUDER, R., AND SURE, Y. SEAL: a framework for developing SEmantic PortALs. In *Proceedings of the 1st international conference on Knowledge capture* (2001), ACM, pp. 155–162.

[194] STONE, L. The Revival of Narrative: Reflections on a New Old History. *Past & Present*, 85 (1979), 3–24.

[195] SUOMINEN, O. *Methods for Building Semantic Portals*. PhD thesis, Aalto University, School of Science, Helsinki, September 2013.

[196] TAMPER, M., LESKINEN, P., IKKALA, E., OKSANEN, A., MÄKELÄ, E., HEINO, E., TUOMINEN, J., KOHO, M., AND HYVÖNEN, E. AATOS – a Configurable Tool for Automatic Annotation. In *Proceedings, Language, Technology and Knowledge (LDK 2017)* (Galway, Ireland, June 2017), Springer, Cham, pp. 276–289.

[197] TEPORA, T. Finnish Civil War 1918. In *1914-1918-online: International Encyclopedia of the First World War*, U. Daniel, P. Gatrell, O. Janz, H. Jones, J. Keene, A. Kramer, and B. Nasson, Eds. Freie Universität Berlin, October 2014.

[198] THOMAS, S., WESSMAN, A., TUOMINEN, J., KOHO, M., IKKALA, E., HYVÖNEN, E., ROHIOLA, V., AND SALMELA, U. SuALT: Collaborative Research Infrastructure for Archaeological Finds and Public Engagement through Linked Open Data. In *Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018), Book of Abstracts* (Helsinki, Finland, March 2018).

[199] TRONCY, R., MALOCHA, B., AND FIALHO, A. T. Linking Events With Media. In *Proceedings of the 6th international conference on semantic systems* (2010), ACM, p. 42.

[200] TUMMARELLO, G., CYGANIAK, R., CATASTA, M., DANIELCZYK, S., DELBRU, R., AND DECKER, S. Sig.ma: Live Views on the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web 8*, 4 (2010), 355–364.

[201] TUNKELANG, D. *Faceted Search*. Synthesis lectures on information concepts, retrieval, and services. Morgan & Claypool Publishers, 2009.

[202] TUOMINEN, J., HYVÖNEN, E., AND LESKINEN, P. Bio CRM: A Data Model for Representing Biographical Data for Prosopographical Research. In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)* (Linz, Austria, 2018), A. Fokkens, S. ter Braake, R. Sluijter, P. Arthur, and E. Wandl-Vogt, Eds., vol. 2119, CEUR Workshop Proceedings, Aachen, Germany, pp. 59–66.

[203] ULRICH, L. T., GASKELL, I., SCHECHNER, S., CARTER, S. A., AND VAN GERBIG, S. *Tangible things: Making history through objects*. Oxford University Press, 2015.

[204] UMBRICH, J., VILLAZÓN-TERRAZAS, B., AND HAUSENBLAS, M. Dataset Dynamics Compendium: A Comparative Study. In *Proceedings of the First International Workshop on Consuming Linked Data (COLD2010)* (Shanghai, China, November 2010), O. Hartig, A. Harth, and J. Sequeda, Eds., vol. 665, CEUR Workshop Proceedings, Aachen, Germany.

[205] UOTILA, M. Tavallisuuden tavoittelua: prosopografia elämäkerrallisen tutkimuksen välineenä. In *Historiallinen elämä: biografia ja historiantutkimus*, H. Hakosalo, S. Jalagin, M. Junila, and H. Kurvinen, Eds. Suomalaisen Kirjallisuuden Seura, 2014, pp. 240–256.

[206] VAN HAGE, W. R., MALAISÉ, V., SEGERS, R., HOLLINK, L., AND SCHREIBER, G. Design and use of the Simple Event Model (SEM). *Journal of Web Semantics 9*, 2 (2011), 128–136.

[207] VAN LEEUWEN, M. H. D., AND MAAS, I. *HISCLASS: A Historical International Social Class Scheme*. Leuven University Press, 2011.

[208] VAN LEEUWEN, M. H. D., MAAS, I., AND MILES, A. *HISCO: Historical International Standard Classification of Occupations*. Leuven University Press, 2002.

[209] VAN NISPEN, A. EHRI Vocabularies and Linked Open Data: An Enrichment? In *Trust and Understanding: the value of metadata in a digitally joined-up world* (2019), R. Depoortere, T. Gheldof, D. Styven, and J. V. D. Eycken, Eds., vol. 106, ABB: Archives et Bibliothèques de Belgique, pp. 117–122. In press.

[210] VAN NISPEN, A., AND JONGMA, L. Holocaust and World War Two Linked Open Data Developments in the Netherlands. *Umanistica Digitale 3*, 4 (2019).

[211] VAN OSSENBRUGGEN, J., AMIN, A., AND HILDEBRAND, M. Why Evaluating Semantic Web Applications is Difficult. In *Proceedings of the Fifth International Workshop on Semantic Web User Interaction (SWUI 2008)* (Florence, Italy, April 2008), D. Degler, mc schraefel, J. Golbeck, A. Bernstein, and L. Rutledge, Eds., vol. 543, CEUR Workshop Proceedings, Aachen, Germany.

[212] VAN VEEN, T., LONIJ, J., AND FABER, W. J. Linking Named Entities in Dutch Historical Newspapers. In *Metadata and Semantics Research* (2016), E. Garoufallou, I. Subirats Coll, A. Stellato, and J. Greenberg, Eds., vol. 672 of *Communications in Computer and Information Science*, Springer International Publishing, Springer, Cham, pp. 205–210.

[213] VELTMAN, K. H. Syntactic and Semantic Interoperability: New Approaches to Knowledge and the Semantic Web. *New Review of Information Networking 7*, 1 (2001), 159–183.

[214] VERBOVEN, K., CARLIER, M., AND DUMOLYN, J. A Short Manual to the Art of Prosopography. In *Prosopography Approaches and Applications. A Handbook*, K. Keats-Rohan, Ed. Unit for Prosopographical Research (Linacre College), 2007, pp. 35–70.

[215] WANG, B., DONG, H., BOEDIHARDJO, A. P., LU, C.-T., YU, H., CHEN, I.-R., AND DAI, J. An Integrated Framework for Spatio-Temporal-Textual Search and Mining. In *Proceedings of the 20th international conference on advances in geographic information systems* (2012), ACM, pp. 570–573.

[216] WARREN, R. Creating specialized ontologies using Wikipedia: The Muninn Experience. In *Proceedings of Wikipedia Academy: Research and Free Knowledge. (WPAC2012)* (Berlin, Germany, June 2012), Wikimedia Deutschland.

[217] WASINSKI, C. On making war possible: Soldiers, strategy, and military grand narrative. *Security Dialogue 42*, 1 (2011), 57–76.

[218] WEINBERG, G. L. *A world at arms: A global history of World War II.* Cambridge University Press, 1995.

[219] WESSMAN, A., THOMAS, S., ROHIOLA, V., KOHO, M., IKKALA, E., TUOMINEN, J., HYVÖNEN, E., KUITUNEN, J., PARVIAINEN, H., AND NIUKKANEN, M. Citizen Science in Archaeology: Developing a Collaborative Web Service for Archaeological Finds in Finland. In *Transforming Heritage Practice in the 21st Century: Contributions from Community Archaeology*, J. Jameson and S. Musteaţă, Eds. Springer, Cham, July 2019, pp. 337–352.

[220] WESSMAN, A., THOMAS, S., ROHIOLA, V., KUITUNEN, J., IKKALA, E., TUOMINEN, J., KOHO, M., AND HYVÖNEN, E. A Citizen Science Approach to Archaeology: Finnish Archaeological Finds Recording Linked Open Database (SuALT). In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN 2019)* (Copenhagen, Denmark, March 2019), C. Navarretta, M. Agirrezabal, and B. Maegaard, Eds., vol. 2364, CEUR Workshop Proceedings, Aachen, Germany, pp. 469–478.

[221] WINER, D. Review of Ontology Based Storytelling Devices. In *Language, Culture, Computation. Computing of the Humanities, Law, and Narratives*, N. Dershowitz and E. Nissan, Eds., vol. 8002 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 2014, pp. 394–405.

[222] YLI-LUUKKO, E. Kirjeisiin kirjoitettu sota: avioparin kirjeenvaihdossa rakentuvat merkitykset vuonna 1944. Master's thesis, University of Jyväskylä, Faculty of Humanities, Department of History and Ethnology, January 2015.

[223] YLIKANGAS, H. *Mitä on historia ja millaista sen tutkiminen*. Art House, Helsinki, 2015.

[224] ZABLITH, F., ANTONIOU, G., D'AQUIN, M., FLOURIS, G., KONDYLAKIS, H., MOTTA, E., PLEXOUSAKIS, D., AND SABOU, M. Ontology evolution: a process-centric survey. *The Knowledge Engineering Review 30*, 1 (2015), 45–75.

[225] ZABLITH, F., SABOU, M., D'AQUIN, M., AND MOTTA, E. Ontology Evolution with Evolva. In *The Semantic Web: Research and Applications* (2009), L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, Eds., Springer Berlin Heidelberg, pp. 908–912.

[226] ZENG, M. L., AND QIN, J. *Metadata*, 2nd ed. Facet Publishing, London, UK, 2016.

[227] ZHAO, J., BIZER, C., GIL, A., MISSIER, P., AND SAHOO, S. Provenance Requirements for the Next Version of RDF. In *Proceedings of the W3C Workshop – RDF Next Steps* (2010), W3C. https://www.w3.org/2009/12/rdf-ws/papers/ws08, [Accessed 12.10.2019].

[228] ZHAO, J., AND HARTIG, O. Towards Interoperable Provenance Publication on the Linked Data Web. In *Proceedings of the 5th Linked Data on the Web (LDOW) Workshop at the World Wide Web Conference (WWW)* (Lyon, France, April 2012), C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, Eds., vol. 937, CEUR Workshop Proceedings, Aachen, Germany.

Publication I

Mikko Koho, Esko Ikkala, Petri Leskinen, Minna Tamper, Jouni Tuominen, and Eero Hyvönen. WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data. Submitted to *Semantic Web – Interoperability, Usability, Applicability: Special Issue on Semantic Web for Cultural Heritage*, October 2019.

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-

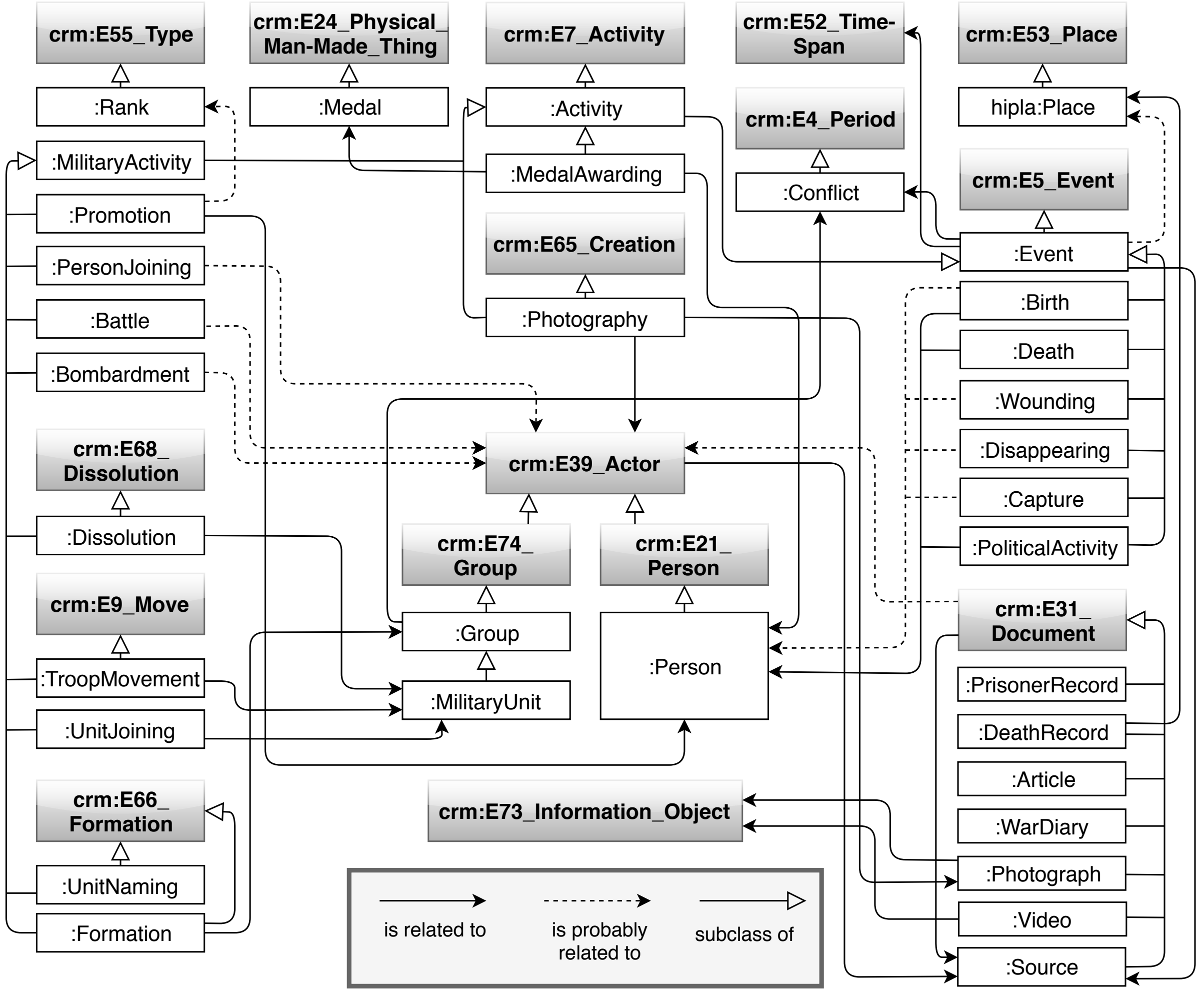


Figure 1. The CRM based WarSampo data model for representing military history as events.

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.

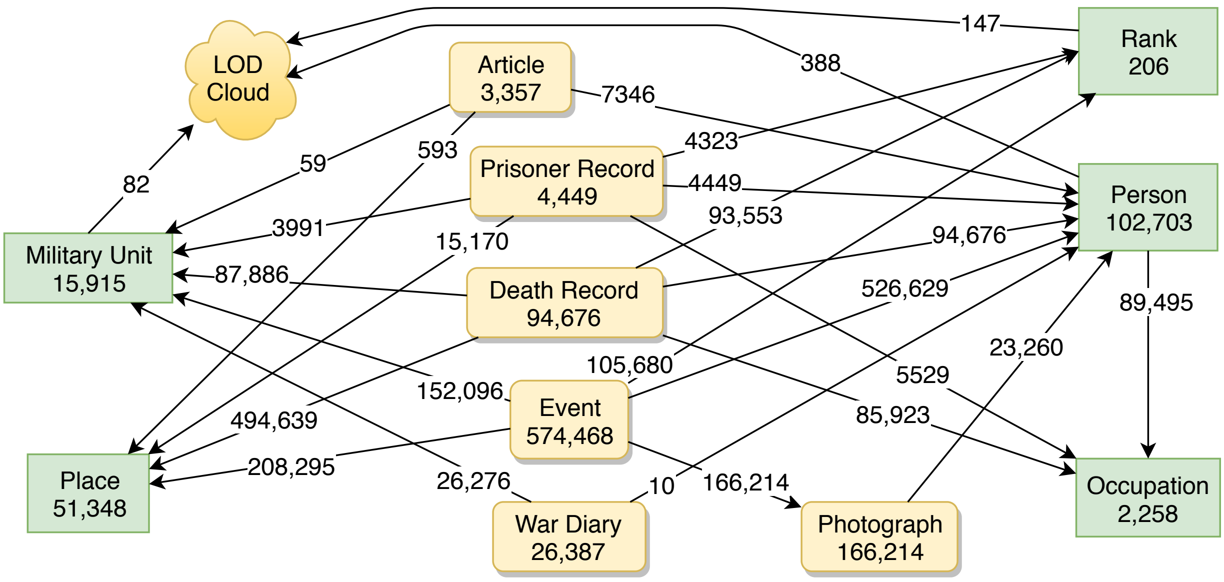


Figure 2. WarSampo core classes with instance counts and linkage between the class instances.

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

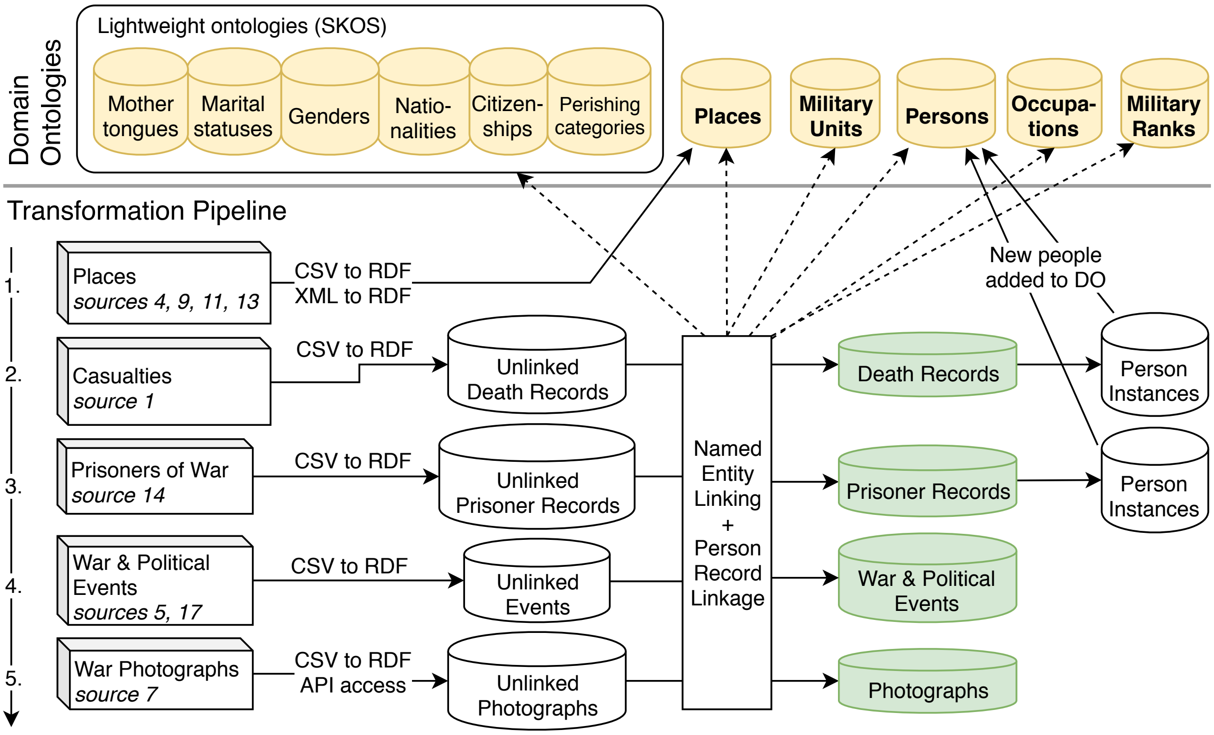


Figure 3. The 5-step WarSampo data transformation process. Dashed arrows represent entity linking, while solid arrows convey data flow.

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 CrossRef 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.

Publication II

Petri Leskinen, Mikko Koho, Erkki Heino, Minna Tamper, Esko Ikkala, Jouni Tuominen, Eetu Mäkelä, and Eero Hyvönen. Modeling and Using an Actor Ontology of Second World War Military Units and Personnel. In *The Semantic Web – ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part II*, Claudia d'Amato, Miriam Fernandez, Valentina Tamma, Freddy Lecue, Philippe Cudré-Mauroux, Juan Sequeda, Christoph Lange, and Jeff Heflin (editors), Information Systems and Applications, incl. Internet/Web, and HCI, volume 10588, pages 280–296, ISBN 9783319682037, Springer, Cham, October 2017.

Modeling and Using an Actor Ontology of Second World War Military Units and Personnel

Petri Leskinen[1], Mikko Koho[1], Erkki Heino[1,2], Minna Tamper[1,2], Esko Ikkala[1], Jouni Tuominen[1,2], Eetu Mäkelä[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. This paper presents a model for representing historical military personnel and army units, based on large datasets about World War II in Finland. The model is in use in WarSampo data service and semantic portal, which has had tens of thousands of distinct visitors. A key challenge is how to represent ontological changes, since the ranks and units of military personnel, as well as the names and structures of army units change rapidly in wars. This leads to serious problems in both search as well as data linking due to ambiguity and homonymy of names. In our solution, actors are represented in terms of the events they participated in, which facilitates disambiguation of personnel and units in different spatio-temporal contexts. The linked data in the WarSampo Linked Open Data cloud and service has ca. 9 million triples, including actor datasets of ca. 100 000 soldiers and ca. 16 100 army units. To test the model in practice, an application for semantic search and recommending based on data linking was created, where the spatio-temporal life stories of individual soldiers can be reassembled dynamically by linking data from different datasets. An evaluation is presented showing promising results in terms of linking precision.

Keywords: Semantic Web, Linked Open Data, Actor Ontology, Digital Humanities, Biographic Representation

1 Introduction

Authority files [18], vocabularies (e.g., ULAN [3]), and actor ontologies (e.g. FOAF [4], REL [5], BIO [6], schema.org [5]) are used for 1) identifying people, groups, and organizations and 2) for representing data about them. They constitute a central resource for cataloging and information management in museums, libraries, and archives, but also a challenge for data linking due to alternative names, homonyms, spelling variations,

[3] `http://www.getty.edu/research/tools/vocabularies/ulan/about.html`

[4] `http://xmlns.com/foaf/spec/`

[5] `http://vocab.org/relationship/`

[6] `http://vocab.org/bio/`

different languages, transliteration rules, and changes in time. Although actor ontologies play an essential part in modeling historical information, there are still very few published scientific articles about the subject.

Historical military units and personnel is a particularly challenging domain for creating an actor ontology: the structures of units are large and change rapidly, different codes can be used for actors in order to confuse the enemies, and people come and go due to the violent actions of war. For example, during the phases of WW2 in Finland (The Winter War, The Continuation War, and The Lapland War) different units have used the same name, and during Winter War in Finland the names of major units were changed just to bluff the enemy. Furthermore, the data about the actors is often incomplete and uncertain, involving lots of "unknown soldiers" of whom little is known.

From a Linked Data viewpoint this poses two major problems: 1) Data linking (based on named entity linking [6,2]) is difficult, because it has to be done in a changing and vague domain specific contexts [7]. For example, to tell whether a mention *captain Smith* and *colonel Smith* can refer to the same person, and to which *Smith* in the first place, data about different Smiths and their ranking history in time is needed. 2) It is difficult to aggregate and enrich data about actors that come from different sources and in different documentary forms, such as death records, diaries, magazine articles, or photographs, and to compile the global biographical history of the actors to the end users [10].

We argue that to address the problems above, a semantically rich spatio-temporal model for representing actors in relation to the events of the war is needed. This paper contributes to the state-of-the-art by presenting such an ontological actor model for historical military units and personnel. The model is in use in end-use application perspectives of the semantic portal WarSampo[7], where the idea is to reassemble automatically the biographical war history of individual soldiers and units. The model enables disambiguation of names in spatio-temporal contexts as well as combining contents from various sources, and publishing them in a harmonized format. The ontology and related data has been published as a Linked Open Data service[8] that can and has been used in digital humanities research and as well for developing online portals. For example, the community portal Sotapolku[9], provided by a commercial company, makes use of the WarSampo actor data.

The work is done as part of the WarSampo project[10], and builds upon our previous publications [12,7,8,10], which focus on the architecture, named entity linking, and end-user views of the application. In contrast, this paper represents the underlying ontology model and dataset regarding army units and people in detail, as well as the actor related application perspectives in use.

The paper is structured as follows: First, ontology model for representing army units, and military personnel, is presented. After this the collecting of WarSampo actor dataset is represented, and a brief look on person and unit perspectives at WarSampo

[7] Sotasampo in Finnish; available with an English GUI at `http://www.sotasampo.fi/en`, but the content is in Finnish.

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

[9] `http://sotapolku.fi`

[10] `http://seco.cs.aalto.fi/projects/sotasampo/en/`

portal is taken. In conclusion, contributions of the work are summarized and some directions for further research are suggested.

2 Use Case and Datasets

The use case for our work is the WarSampo semantic portal[11] [10]. It provides the end user with richly interlinked data about the WW2 in Finland via application perspectives in the Sampo model [9]. An illustration of the WarSampo datasets is represented in Figure 1. In total, the WarSampo data cloud contains data of more than a dozen different types (e.g. casualty data, photographs, events, war diaries, and historical maps) from an even larger pool of sources (e.g. the National Archives, the Defense Forces, and scanned books, from which part of the data has been extracted semi-automatically).

The actor dataset contains ca. 100 000 soldiers, and ca. 16 100 army units. The data is enriched with ca. 488 000 links from events to actors. Actors have furthermore been linked to external resources in the LOD cloud databases on the Web.

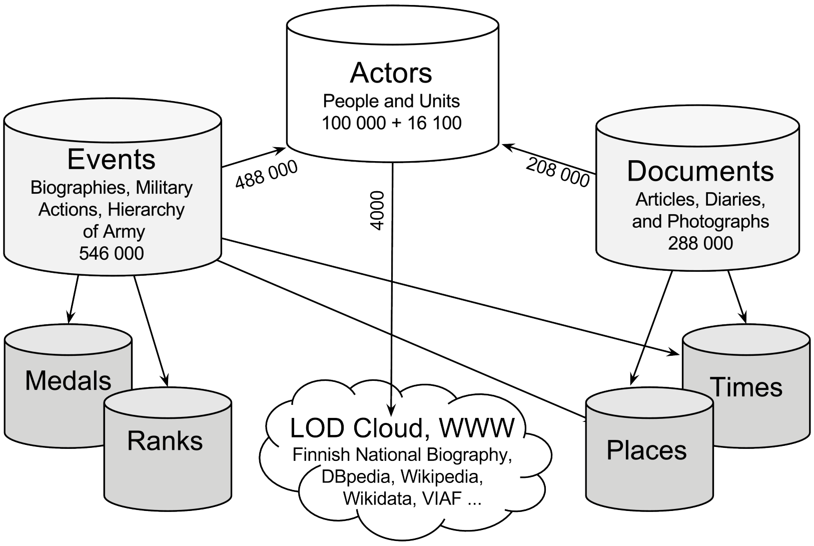

Actors
People and Units
100 000 + 16 100
Events
Biographies, Military Actions, Hierarchy of Army
546 000
488 000
4000
208 000
Documents
Articles, Diaries, and Photographs
288 000
Medals
Ranks
LOD Cloud, WWW
Finnish National Biography, DBpedia, Wikipedia, Wikidata, VIAF ...
Places
Times

Fig. 1. Linkage in the actor-event based dataset

3 Actor Ontology Model

The ontology of actors is based on the CIDOC CRM[12] [4] model, where the resources of actors are essentially described in terms of the spatio-temporal events they participate

[11] http://sotasampo.fi/en/

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

in. An event represents any change of status that divides the timeline into periods before and after the event. Using the actor-event-model facilitates reconstructing the status of an actor at a specified moment. One main reason for adapting the model is that the information regarding a single actor varies a lot in both form and amount; in some cases we may have access to a very detailed description of the actor's biography, in some other cases only sparse pieces of information exist. All this data can be harmonized into a sequence of events. The applied actor-event-model also allows us to easily add new event types to the schema and new events the to database.

Table 1. Namespaces and prefixes used in actor ontologies

Namespace	**Prefix**
http://ldf.fi/schema/warsa/	:
http://www.cidoc-crm.org/cidoc-crm/	crm:
http://purl.org/dc/elements/1.1/	dc:
http://purl.org/dc/terms/	dct:
http://xmlns.com/foaf/0.1/	foaf:
http://rdf.muninn-project.org/ontologies/organization#	mil:
http://www.w3.org/2002/07/owl#	owl:
http://www.w3.org/1999/02/22-rdf-syntax-ns#	rdf:
http://www.w3.org/2004/02/skos/core#	skos:

Schema of the ontology is illustrated in Figure 2. The schema is available at `http://ldf.fi/schema/warsa`, the namespaces and prefixes in use are listed in Table 1. The actor superclass **crm:E39_Actor**[13] is shown at center on the top. There is one subclass for people, and two for groups. For various types of events there are 19 classes with superclass **:Event**[14].

The biographical representation of a person was modeled with events of birth (**:Birth**), and death (**:Death**), and his military career with events like promotion (**:Promotion**), serving in an army unit (**:UnitJoining**), participating in battles (**:Battle**), or getting awarded with a medal of honor (**:MedalAwarding**). Furthermore, there are classes for getting wounded (**:Wounding**) or disappearing (**:Disappearing**), which represent the data fields in Casualties database. The schema includes supporting classes for representing military ranks, war diary entries, medals of honor, documentation, and data sources.

Example of a person resource[15] (**:Person**) is shown in Table 2. The principle is to represent only constant information in a person resource; it has full name as a primary

[13] `http://www.cidoc-crm.org/cidoc-entities/e39-actor`

[14] `http://www.cidoc-crm.org/cidoc-entities/e5-event`

[15] `http://ldf.fi/warsa/actors/person_294.ttl`

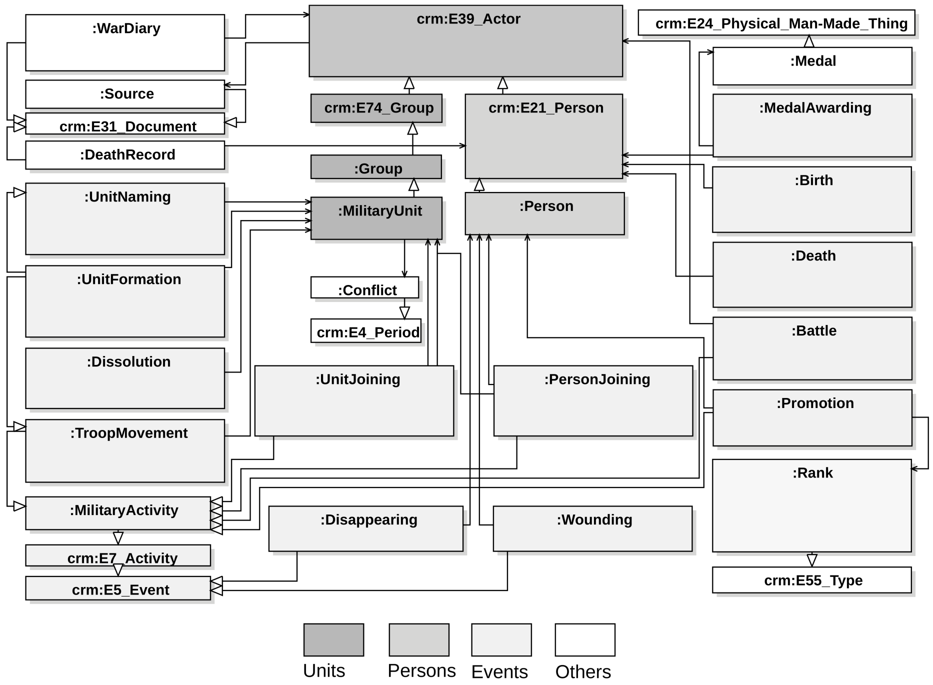


Fig. 2. Ontology schema of actors and events

title, and the family and first names as separate fields. Property **owl:sameAs** links to a corresponding resource in external databases, and **foaf:page** to external web pages.

Examples of related events are shown in Table 3. During the war, the person in example has been promoted from lieutenant first to captain and finally to major. When the Winter War started in 1939 he served as a commander in an air force squadron, and shot down an enemy aircraft soon after.

In literature military personnel are ofter referred using a combination of current military rank and family name (e.g. *Captain Karhunen* or *Colonel Talvela*). So, to describe a person in detail, an ontology of military ranks was needed. The rank ontology is based on Muninn Military Ontology [19]. The hierarchy of ranks was constructed by interlinking the instances to equal and lower ranks. The **:Rank** instances in the datasets were enriched with additional information (e.g. countries or service branches in which the rank has been used, or categories like officer or non-commissioned officer). Event **:Promotion** was used to attach a rank to a person. Due to the variations in the amount of available data, a promotion event was created in all cases, even if a person is known to have only a single rank with no specific date of promotion.

An example of RDF resource of a military unit is shown in Table 4, the resource is also available in Turtle format[16]. Just like in the case of a person, the properties describe only constant information like unit's preferable name and abbreviation, descrip-

[16] `http://ldf.fi/warsa/actors/actor_459`

Table 2. Properties of a resource describing pilot Jorma Karhunen

Property	**RDF identifier**	**Value**
Primary title	skos:prefLabel	”Jorma (Joppe) Karhunen”
Family Name	foaf:familyName	”Karhunen”
First name (Nickname)	foaf:firstName	”Jorma (Joppe)”
Text description	dc:description	“Jorma Karhunen was a Finnish Air ...”@en
External LOD-links	owl:sameAs	http://dbpedia.org/resource/Jorma_Karhunen
		http://wikidata.org/entity/Q5482501
Related websites	foaf:page	https://en.wikipedia.org/wiki/Jorma_Karhunen
		www.mannerheim-ristinritarit.fi/ritarit?xmid=38

tion, conflicts participated in, and links to LOD cloud resources. The events (Table 5) describe the unit's position in the army hierarchy and the involved military activities. The lifespan of a unit spans from its formation **:UnitFormation** to dissolution **:Dissolution**. The changes of the unit name were modeled as **:UnitNaming** events. Also the army hierarchy, including the temporal changes made in it, was modeled using the event schema: the hierarchy was represented as a tree graph where the army units are the nodes and the events of joining into a superior unit **:UnitJoining** form the edges. The events also included the military activities taken (e.g. movements **:TroopMovement** and battles **:Battle**). The event **:PersonJoining** was used to combine a person to the unit, in which he has served. The event could also announce a role in the unit (e.g. being a commander or a squadron pilot).

4 Warsampo Actor Data

Currently the actor dataset contains ca. 100 000 people. The data has been collected from various sources: lists of generals and commanders, lists of recipients of honorary medals, the Casualties database[17], Finnish National Biography[18], photographers mentioned in Finnish Wartime Photograph Archive[19], Wikidata[20], and Wikipedia. Besides military personnel, an extract of 580 Finnish or foreign civilians from the National Biography database and Wikidata was included. This set consisted of people with political or cultural significance.

The unit dataset consists of over 16 100 Finnish wartime units, including Land Forces, Air Forces, Navy, Medical Corps, stations of Anti-Aircraft Warfare and Airwarning, Finnish White Guard, and Foreign Volunteer Corps. At this stage Soviet and

[17] kronos.narc.fi/menehtyneet/

[18] http://www.kansallisbiografia.fi/english/

[19] http://sa-kuva.fi/neo?tem=webneoeng

[20] https://www.wikidata.org/

Table 3. Examples of events describing pilot Jorma Karhunen

Event description / Resource URI	RDF class	date
Born at Pyhäjärvi http://ldf.fi/warsa/events/birth_294.ttl	:Birth	1913-03-17
Serving as a squadron commander in 24th Fighter Squadron http://ldf.fi/warsa/events/joining_294_459.ttl	:PersonJoining	1939-11-30
Aerial victory in Tainionkoski: enemy SB-2 shot down http://ldf.fi/warsa/events/event_lv2408.ttl	:Battle	1939-12-01
Promotion to captain http://ldf.fi/warsa/events/kapteeni_294.ttl	:Promotion	1941-08-04
Photograph of capt. Karhunen with his dog Becky Brown http://ldf.fi/warsa/photographs/sakuva_7265.ttl	:Photography	1942-06-01
Awarded with the Mannerheim Cross of Liberty http://ldf.fi/warsa/medals/medal_83_294.ttl	:MedalAwarding	1942-09-08
Died at Tampere http://ldf.fi/warsa/events/death_294.ttl	:Death	2002-01-18

German troops were excluded. The main sources of information have been the War Diaries, Army Postal Code list[21], and Organization Cards, all of which provided the information as datasheets in CSV format.

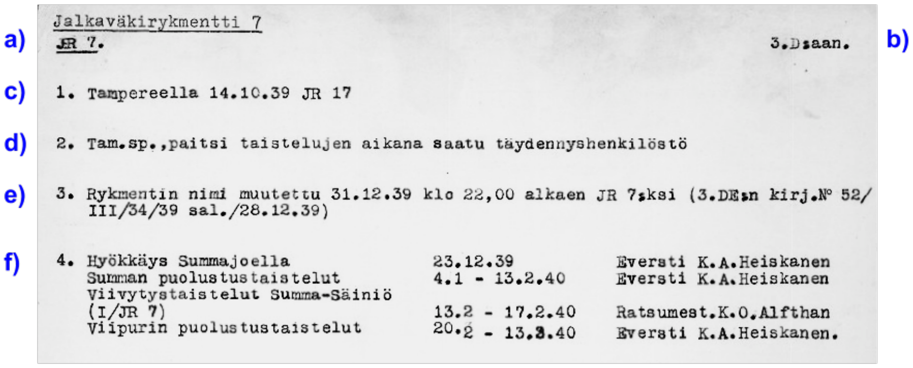
In general, the method to produce the data depended on the format of data source. The biographies of the National Biography and the Casualties Database had been transformed into LOD in our earlier projects, and therefore the information extraction process was to convert the existing data into new actor entries and relating events. Transformation was mostly done by using specific SPARQL construct queries. More than 95 000 entries were generated from the Casualty Database to actor dataset [12].

The organization cards (Figure 3) were written by Finnish Defense Forces shortly after the WW2. The cards contain the major part of units in Finnish Army, unfortunately not those of Navy and Air Force. An example of organization card is shown in Figure 3. The proper name and abbreviation of the unit is shown at the upper left corner (a), in this case *Jalkaväkirykmentti 7* (7th Infantry Regiment), abbreviated as *JR 7*. The regiment has been part of *3. divisioona* (3rd Division), which is told at the upper right corner (b). The card provides further information about the foundation (c) and the military district (d) of the unit. Changes considering the unit, like different names, are shown at part (e). During the Winter War *JR 7* participated in four battles (f). The three columns on each line show the location or a short description of the battle, battle's duration, and the name of the commanding officer.

[21] http://www.arkisto.fi/uploads/Aineistot/kopsa[1].pdf

Table 4. Properties of a resource describing 24th Fighter Squadron

Property	RDF identifier	Value
preferred label	skos:prefLabel	"Lentolaivue 24"
preferred abbreviation	skos:altLabel	"LLv 24"
description	dc:description	"No. 24 Squadron was a fighter ..."@en
conflict	:hasConflict	wcf:WinterWar, wcf:ContinuationWar, ...
Army postal code	:covernumber	"8523", "8524", "8567"
unit category	:hasUnitCategory	"Flying Regiments and Squadrons"
external LOD-links	owl:sameAs	https://www.wikidata.org/wiki/Q4356342
related websites	foaf:page	https://fi.wikipedia.org/wiki/Lentolaivue_24



```
Jalkaväkirykmentti 7
JR 7.                                                          3.Dsaan.

1. Tampereella 14.10.39 JR 17

2. Tam.sp.,paitsi taistelujen aikana saatu täydennyshenkilöstö

3. Rykmentin nimi muutettu 31.12.39 klo 22,00 alkaen JR 7sksi (3.DEsn kirj.№ 52/
   III/34/39 sal./28.12.39)

4. Hyökkäys Summajoella              23.12.39          Eversti K.A.Heiskanen
   Summan puolustustaistelut         4.1 - 13.2.40     Eversti K.A.Heiskanen
   Viivytystaistelut Summa-Säiniö
   (I/JR 7)                          13.2 - 17.2.40    Ratsumest.K.O.Alfthan
   Viipurin puolustustaistelut       20.2 - 13.3.40    Eversti K.A.Heiskanen.
```

Fig. 3. Information on an Organization Card

The organization cards were provided as scanned booklets in PDF format, and converting to RDF had several steps. Firstly each page in PDF booklet was written as an individual PNG image. Images were preprocessed by adjusting the contrast and image rotation, and removing the compression artifacts. Next an *Optical character recognition (OCR)* process was applied. The resulting text was however very erroneous, and plenty of post-processing was required. The structured format of the cards, and the recurring use of military terms in the vocabulary however eased the automated error fixing. From the resulting text, the fields a-f (in Figure 3) were extracted, and converted into RDF. The produced resources consisted of military units (**:MilitaryUnit**), their commanders (**:Person**) with ranks (**:Promotion**), and events like unit formations (**:UnitFormation**), joinings of units (**:UnitJoining**), movements (**:TroopMovement**), renamings (**:UnitNaming**), and battles (**:Battle**).

Although the Wikipedia may not be considered as the most reliable source of information, it provided a way to connect data with external LOD cloud databases Wikidata,

Table 5. Examples of events describing 24th Fighter Squadron

Event description / Resource URI	RDF class	date
Troop founded as 24th Squadron (abbrev. LLv 24) http://ldf.fi/warsa/events/formation_971.ttl	:Formation	1934-10-10
Troop Movement to Immola Air Base http://ldf.fi/warsa/events/concentration_491.ttl	:TroopMovement	1939-10-12
Aerial victory in Tainionkoski: enemy SB-2 shot down http://ldf.fi/warsa/events/event_lv2408.ttl	:Battle	1939-12-01
Being part of Flying Regiment 2 http://ldf.fi/warsa/events/joining_458.ttl	:UnitJoining	1940-01-10
Written War Diary document http://ldf.fi/warsa/diaries/diary_c26701.ttl	:WarDiary	1941-06-19– 1941-09-02
Changing the name to 24th Fighter Squadron (HLeLv 24) http://ldf.fi/warsa/events/form_459.ttl	:UnitNaming	1944-02-14

DBpedia[22], and VIAF[23]. The material regarding personnel was widely available, but for units, specially those of Finnish Army during the WW2, the information was sparse. Information was extracted from Wikipedia pages of e.g. Finnish high-ranking officers, politicians, wartime casualties, and foreign volunteers. The pages of Wikipedia follow a structured layout which facilitated extracting the information. In case of military units, detailed information for events like unit foundation, troop movements, battles, and for names of commanding officers were available. In total 2500 people and 480 units with 5000 events were generated from corresponding Wikipedia pages.

Characteristic sentences picked from Wikipedia were descriptions like *"1st Artillery Group was founded in Pori with Captain Paavo Suominen as the first commander"*, *"10th July 1941 Regiment was moved to Kitee, from where it begun attacking towards Lake Ladoga"*, or *"Regiment participated in the occupation of Prääsä September 7–8, 1941"*. Each sentence was converted to an event, and the named entities of personnel, places and dates were recognized and linked to database resources. The data retrieval was done using Python scripts utilizing MediaWiki API[24], and Wikipedia API for Python[25]. Entity linking was done with ARPA service[13].

The datasets of conflicts, war diaries, medals, and ranks are in separate graphs. Conflicts[26] contain four main periods of WW2 in Finland. The War Diary graph[27] has

[22] `http://wiki.dbpedia.org/`

[23] `https://viaf.org/`

[24] `https://en.wikipedia.org/w/api.php`

[25] `https://pypi.python.org/pypi/wikipedia`

[26] See, e.g., `http://ldf.fi/warsa/conflicts/LaplandWar`

[27] See, e.g., `http://digi.narc.fi/digi/hae_ay.ka?sartun=319.SARK`

26 400 entries. There are 200 medal types[28] and 200 rank entries[29]. The data includes ranks used by the Finnish Military with most common German and Soviet ranks, among with some civil titles (e.g. the ones used by women's voluntary association *Lotta Svärd*). [10]

5 The WarSampo portal

The perspectives at WarSampo portal[30] visualize the linkage between the various datasets (e.g. military unit, personnel, casualties, events, places) etc [10,8]. WarSampo portal is a Rich Internet Application (RIA), where all functionality is implemented on the client side using JavaScript with AngularJS framework, only data is fetched from the server side SPARQL endpoints.

5.1 The Person Perspective

The WarSampo person perspective application[31] is illustrated in Fig. 4. A 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 query results are shown in the text field (b) below the input. After making a selection, information about the person is shown at the center top of the page (c). The tabs (d) allow the user to switch between this information page or a map-timeline application. In the example case, the page shows description of the person (e), photograph gallery (f), lists linking to related events (g), military units (h), battles (i), ranks (j), medals (k), related people (l), places (m), Wikipedia page (n), related Kansa Taisteli magazine articles (o), and a Finnish National Biography widget (p).

As an example of SPARQL query, the query fetching related people[32] defines a similarity measure between two people. The more events, medals, units, and the higher ranks the two share in common, the higher the similarity gets. The list of related people (l) shows the results sorted in descending order.

WarSampo military unit perspective application[33] is illustrated in Figure 5. In a typical use case someone searches for information about an army unit, where perhaps an elder relative has served during the wartime. On the left there is an input field (a) for a search by unit's name. The matching results are shown in the text field (b) below the input. The map (c) depicts the known locations of the unit. The heatmap shows the casualties of the unit, and the timeline (d) the events (e), e.g. dates of unit foundations, troop movements, and durations of fought battles. On the right there are unit names and abbreviations (f), description (g), and a collection of related photographs (h). Three lists of related units are shown: larger groups in which the unit has been as a member (i), subdivisions being parts of the unit (j), and units at the same hierarchical level (k).

[28] See, e.g., `http://ldf.fi/warsa/medals/medal_83`

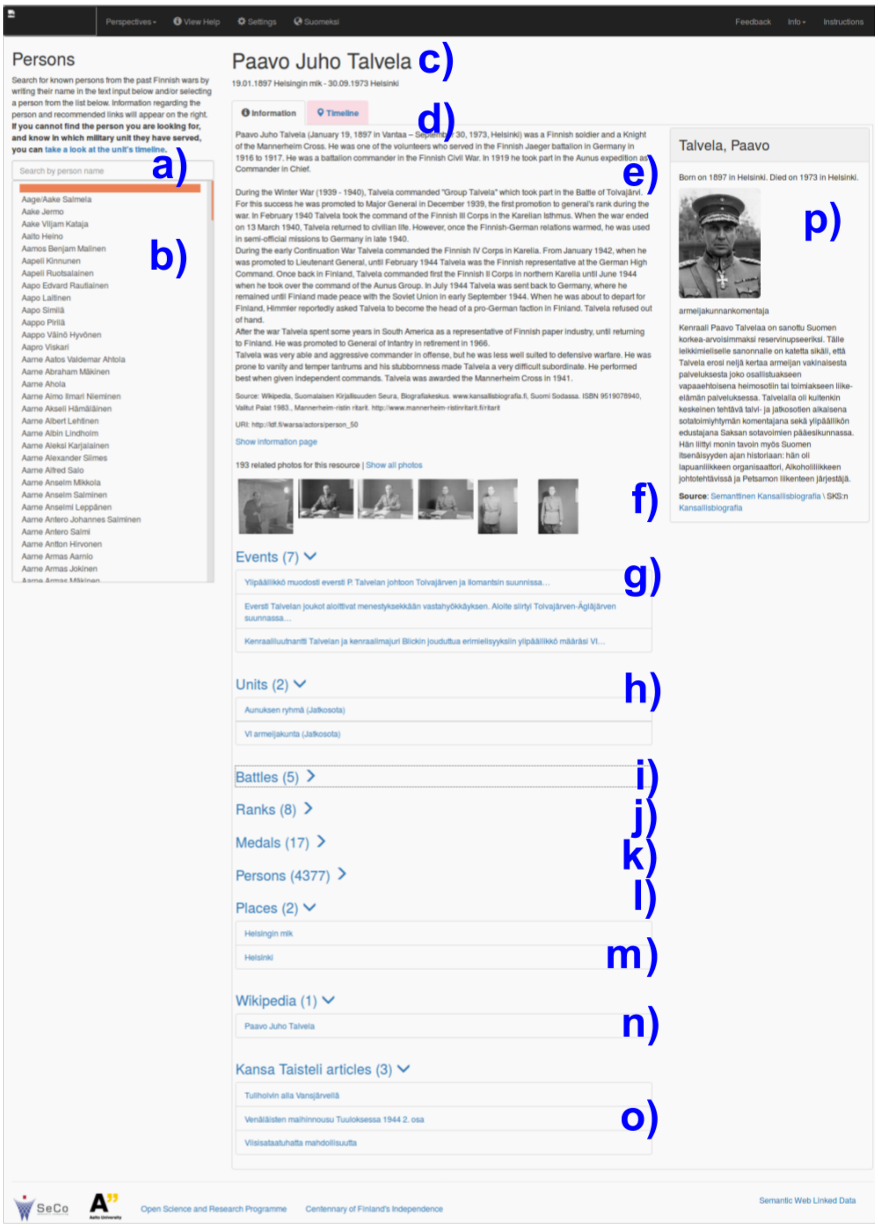
[29] See, e.g., `http://ldf.fi/warsa/actors/ranks/Majuri`

[30] `http://www.sotasampo.fi/en/`

[31] `http://www.sotasampo.fi/en/persons`

[32] `http://yasgui.org/short/B1w2O71gb`

[33] `http://sotasampo.fi/en/units`



Persons

Search for known persons from the past Finnish wars by writing their name in the text input below and/or selecting a person from the list below. Information regarding the person and recommended links will appear on the right. **If you cannot find the person you are looking for, and know in which military unit they have served, you can take a look at the unit's timeline.**

Search by person name

Aage/Aake Salmela
Aake Jermo
Aake Viljam Kataja
Aalto Heino
Aamos Benjam Malinen
Aapeli Kinnunen
Aapeli Ruotsalainen
Aapo Edvard Rautiainen
Aapo Laitinen
Aapo Similä
Aappo Pirilä
Aappo Väinö Hyvönen
Aapro Viskari
Aarne Aatos Valdemar Ahtola
Aarne Abraham Mäkinen
Aarne Ahola
Aarne Aimo Ilmari Nieminen
Aarne Akseli Hämäläinen
Aarne Albert Lehtinen
Aarne Albin Lindholm
Aarne Aleksi Karjalainen
Aarne Alexander Siimes
Aarne Alfred Salo
Aarne Anselm Mikkola
Aarne Anselm Salminen
Aarne Anselmi Leppänen
Aarne Antero Johannes Salminen
Aarne Antero Salmi
Aarne Antton Hirvonen
Aarne Armas Aarnio
Aarne Armas Jokinen
Aarne Armas Mäkinen

Paavo Juho Talvela

19.01.1897 Helsingin mlk - 30.09.1973 Helsinki

Paavo Juho Talvela (January 19, 1897 in Vantaa – September 30, 1973, Helsinki) was a Finnish soldier and a Knight of the Mannerheim Cross. He was one of the volunteers who served in the Finnish Jaeger battalion in Germany in 1916 to 1917. He was a battalion commander in the Finnish Civil War. In 1919 he took part in the Aunus expedition as Commander in Chief.

During the Winter War (1939 - 1940), Talvela commanded "Group Talvela" which took part in the Battle of Tolvajärvi. For this success he was promoted to Major General in December 1939, the first promotion to general's rank during the war. In February 1940 Talvela took the command of the Finnish III Corps in the Karelian Isthmus. When the war ended on 13 March 1940, Talvela returned to civilian life. However, once the Finnish-German relations warmed, he was used in semi-official missions to Germany in late 1940.
During the early Continuation War Talvela commanded the Finnish IV Corps in Karelia. From January 1942, when he was promoted to Lieutenant General, until February 1944 Talvela was the Finnish representative at the German High Command. Once back in Finland, Talvela commanded first the Finnish II Corps in northern Karelia until June 1944 when he took over the command of the Aunus Group. In July 1944 Talvela was sent back to Germany, where he remained until Finland made peace with the Soviet Union in early September 1944. When he was about to depart for Finland, Himmler reportedly asked Talvela to become the head of a pro-German faction in Finland. Talvela refused out of hand.
After the war Talvela spent some years in South America as a representative of Finnish paper industry, until returning to Finland. He was promoted to General of Infantry in retirement in 1966.
Talvela was very able and aggressive commander in offense, but he was less well suited to defensive warfare. He was prone to vanity and temper tantrums and his stubbornness made Talvela a very difficult subordinate. He performed best when given independent commands. Talvela was awarded the Mannerheim Cross in 1941.

Source: Wikipedia, Suomalaisen Kirjallisuuden Seura, Biografiakeskus. www.kansallisbiografia.fi, Suomi Sodassa. ISBN 9519078940, Valitut Palat 1983., Mannerheim-ristin ritarit. http://www.mannerheim-ristinritarit.fi/ritarit

URI: http://ldf.fi/warsa/actors/person_50

Show information page

193 related photos for this resource | Show all photos

Events (7)

Ylipäällikkö muodosti eversti P. Talvelan johtoon Tolvajärven ja Ilomantsin suunnissa...

Eversti Talvelan joukot aloittivat menestyksekkään vastahyökkäyksen. Aloite siirtyi Tolvajärven-Äglajärven suunnassa...

Kenraaliluutnantti Talvelan ja kenraalimajuri Blickin jouduttua erimielisyyksiin ylipäällikkö määräsi VI...

Units (2)

Aunuksen ryhmä (Jatkosota)

VI armeijakunta (Jatkosota)

Battles (5)

Ranks (8)

Medals (17)

Persons (4377)

Places (2)

Helsingin mlk

Helsinki

Wikipedia (1)

Paavo Juho Talvela

Kansa Taisteli articles (3)

Tulihoivin alla Vansjärvellä

Venäläisten maihinnousu Tuuloksessa 1944 2. osa

Viisisataatuhatta mahdollisuutta

Talvela, Paavo

Born on 1897 in Helsinki. Died on 1973 in Helsinki.

armeijakunnankomentaja

Kenraali Paavo Talvelaa on sanottu Suomen korkea-arvoisimmaksi reservinupseeriksi. Tälle leikkimieliselle sanonnalle on katetta sikäli, että Talvela erosi neljä kertaa armeijan vakinaisesta palveluksesta joko osallistuakseen vapaaehtoisena heimosotiin tai toimiakseen liike-elämän palveluksessa. Talvelalla oli kuitenkin keskeinen tehtävä talvi- ja jatkosotien aikaisena sotatoimiyhtymän komentajana sekä ylipäällikön edustajana Saksan sotavoimien pääesikunnassa. Hän liittyi monin tavoin myös Suomen itsenäisyyden ajan historiaan: hän oli lapuanliikkeen organisaattori, Alkoholiliikkeen johtotehtävissä ja Petsamon liikenteen järjestäjä.

Source: Semanttinen Kansallisbiografia \ SKS:n Kansallisbiografia

SeCo

Aalto University

Open Science and Research Programme

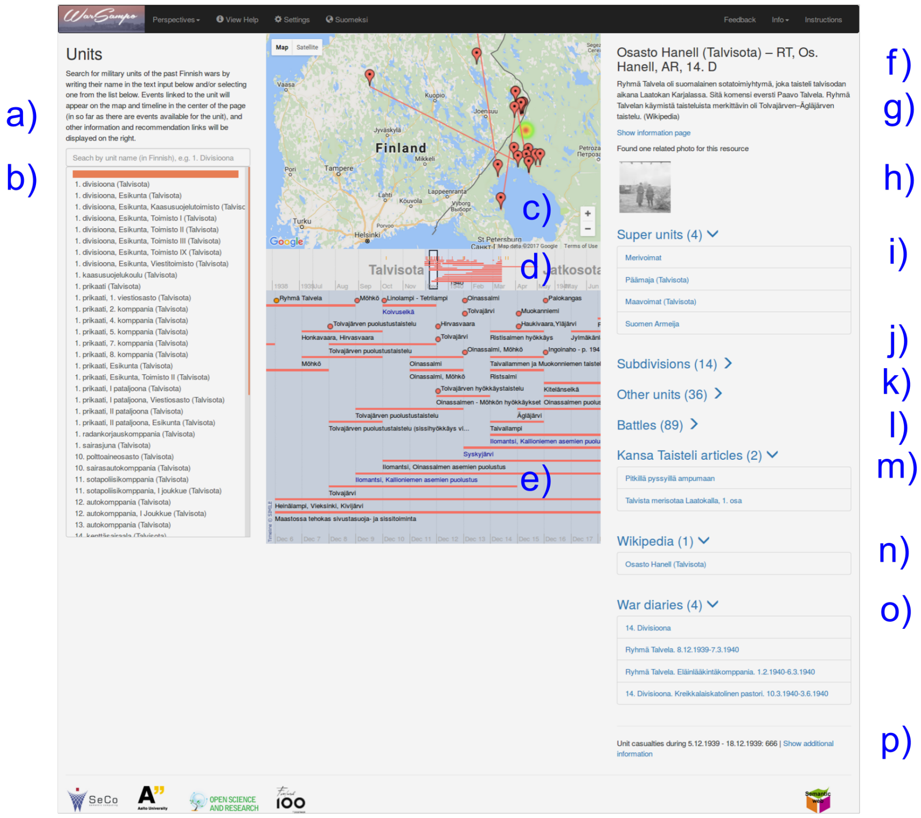
Centennary of Finland's Independence

Semantic Web Linked Data

Fig. 4. Information on Person Perspective

Below there are fields for related battles (l), links to Kansa Taisteli magazine articles (m), Wikipedia page (n), and War Diaries (o). The number of casualties during the specified time is shown at the bottom of page (p).

5.2 The Military Unit Perspective



Units

Search for military units of the past Finnish wars by writing their name in the text input below and/or selecting one from the list below. Events linked to the unit will appear on the map and timeline in the center of the page (in so far as there are events available for the unit), and other information and recommendation links will be displayed on the right.

Seach by unit name (in Finnish), e.g. 1. Divisioona

1. divisioona (Talvisota)
1. divisioona, Esikunta (Talvisota)
1. divisioona, Esikunta, Kaasusuojelutoimisto (Talvisc
1. divisioona, Esikunta, Toimisto I (Talvisota)
1. divisioona, Esikunta, Toimisto II (Talvisota)
1. divisioona, Esikunta, Toimisto III (Talvisota)
1. divisioona, Esikunta, Toimisto IX (Talvisota)
1. divisioona, Esikunta, Viestitoimisto (Talvisota)
1. kaasusuojelukoulu (Talvisota)
1. prikaati (Talvisota)
1. prikaati, 1. viestiosasto (Talvisota)
1. prikaati, 2. komppania (Talvisota)
1. prikaati, 4. komppania (Talvisota)
1. prikaati, 5. komppania (Talvisota)
1. prikaati, 7. komppania (Talvisota)
1. prikaati, 8. komppania (Talvisota)
1. prikaati, Esikunta (Talvisota)
1. prikaati, Esikunta, Toimisto II (Talvisota)
1. prikaati, I pataljoona (Talvisota)
1. prikaati, I pataljoona, Viestiosasto (Talvisota)
1. prikaati, II pataljoona (Talvisota)
1. prikaati, II pataljoona, Esikunta (Talvisota)
1. radankorjauskomppania (Talvisota)
1. sairasjuna (Talvisota)
10. polttoaineosasto (Talvisota)
10. sairasautokomppania (Talvisota)
11. sotapoliisikomppania (Talvisota)
11. sotapoliisikomppania, I joukkue (Talvisota)
12. autokomppania (Talvisota)
12. autokomppania, I Joukkue (Talvisota)
13. autokomppania (Talvisota)
14. kenttäsairaala (Talvisota)

Osasto Hanell (Talvisota) – RT, Os. Hanell, AR, 14. D

Ryhmä Talvela oli suomalainen sotatoimiyhtymä, joka taisteli talvisodan aikana Laatokan Karjalassa. Sitä komensi eversti Paavo Talvela. Ryhmä Talvelan käymistä taisteluista merkittävin oli Tolvajärven–Äglajärven taistelu. (Wikipedia)

Show information page

Found one related photo for this resource

Super units (4)
- Merivoimat
- Päämaja (Talvisota)
- Maavoimat (Talvisota)
- Suomen Armeija

Subdivisions (14)

Other units (36)

Battles (89)

Kansa Taisteli articles (2)
- Pitkillä pyssyillä ampumaan
- Talvista merisotaa Laatokalla, 1. osa

Wikipedia (1)
- Osasto Hanell (Talvisota)

War diaries (4)
- 14. Divisioona
- Ryhmä Talvela. 8.12.1939-7.3.1940
- Ryhmä Talvela. Eläinlääkintäkomppania. 1.2.1940-6.3.1940
- 14. Divisioona. Kreikkalaiskatolinen pastori. 10.3.1940-3.6.1940

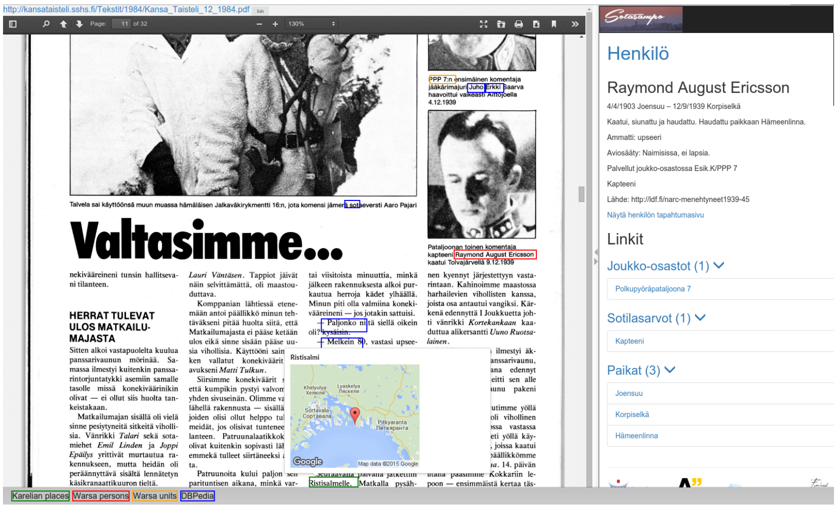
Unit casualties during 5.12.1939 - 18.12.1939: 666 | Show additional information

Fig. 5. Information on Unit Perspective

5.3 The Kansa Taisteli Magazine Perspective

Kansa Taisteli is a magazine published by Sanoma Ltd and Sotamuisto association between 1957 and 1986. The magazine articles cover the memoirs of WW2 from the point of view of Finnish military personnel and civilians. The articles contain mentions of people, military units, and places. From these the military units and personnel have been linked to Actor ontology. The magazine perspective[34] can be used for searching and browsing articles relating to WW2. Military units and personnel are used as separate facets to search for articles. In addition, writers have been linked to Actor ontology as well.

[34] http://sotasampo.fi/en/articles



http://kansataisteli.sshs.fi/Tekstit/1984/Kansa_Taisteli_12_1984.pdf

PPP 7:n ensimäinen komentaja jääkärimajuri Juho Erkki Saarva haavoittui vaikeasti Aittojoella 4.12.1939

Talvela sai käyttöönsä muun muassa hämäläisen Jalkaväkirykmentti 16:n, jota komensi jämerä sotaeversti Aaro Pajari

Pataljoonan toinen komentaja kapteeni Raymond August Ericsson kaatui Tolvajärvellä 9.12.1939

Valtasimme...

nekivääreineni tunsin hallitsevani tilanteen.

HERRAT TULEVAT ULOS MATKAILUMAJASTA

Sitten alkoi vastapuolelta kuulua panssarivaunun mörinää. Samassa ilmestyi kuitenkin panssarintorjuntatykki asemiin samalle tasolle missä konekiväärinikin olivat — ei ollut siis huolta tankeistakaan.

Matkailumajan sisällä oli vielä sinne pesiytyneitä sitkeitä vihollisia. Vänrikki *Talari* sekä sotamiehet *Emil Linden* ja *Joppi Epäilys* yrittivät murtautua rakennukseen, mutta heidän oli peräännyttävä sisältä lennätetyn käsikranaattikuuron tieltä.

Lauri Väntäsen. Tappiot jäivät näin selvittämättä, oli maastouduttava.

Komppanian lähtiessä etenemään antoi päällikkö minun tehtäväkseni pitää huolta siitä, että Matkailumajasta ei pääse ketään ulos eikä sinne sisään pääse uusia vihollisia. Käyttööni sain
ken vallatut konekiväärit
avukseni *Matti Tulkun*.
Siirsimme konekiväärit s
että kumpikin pystyi valvom
yhden sivuseinän. Olimme va
lähellä rakennusta — sisällä
joiden olisi ollut helppo tul
meidät, jos olisivat tuntenee
lanteen. Patruunalaatikkok
olivat kuitenkin sopivasti läh
emmekä tulleet siirtäneeksi
ta.
Patruunoita kului paljon sen parituntisen aikana, minkä var-

tai viisitoista minuuttia, minkä jälkeen rakennuksesta alkoi purkautua herroja kädet ylhäällä. Minun piti olla valmiina konekivääreineni — jos jotakin sattuisi.

— Paljonko niitä siellä oikein oli? kysäisin.

— Melkein 80, vastasi upsee-

Seuraavana päivänä jatkettiin
Ristisalmelle. Matkalla pysäh-

nen kyennyt järjestettyyn vastarintaan. Kahinoimme maastossa harhailevien vihollisten kanssa, joista osa antautui vangiksi. Kärkenä edennyttä I Joukkuetta johti vänrikki *Kortekankaan* kaaduttua alikersantti *Uuno Ruotsalainen*.

ilmestyi äk-
anssarivaunu,
ana edennyt
eitti sen alle
uunu pakeni

utimme yöllä
oli vihollinen
ossa vastassa
eti yöllä käy-
, joissa kaatui
päällikkömme
a. 14. päivän
litana pääsimme Kokkariin le-
poon — ensimmäistä kertaa täs-

Karelian places Warsa persons Warsa units DBPedia

Henkilö

Raymond August Ericsson

4/4/1903 Joensuu – 12/9/1939 Korpiselkä

Kaatui, siunattu ja haudattu. Haudattu paikkaan Hämeenlinna.

Ammatti: upseeri

Aviosääty: Naimisissa, ei lapsia.

Palvellut joukko-osastossa Esik.K/PPP 7

Kapteeni

Lähde: http://ldf.fi/narc-menehtyneet1939-45

Näytä henkilön tapahtumasivu

Linkit

Joukko-osastot (1)

Polkupyöräpataljoona 7

Sotilasarvot (1)

Kapteeni

Paikat (3)

Joensuu

Korpiselkä

Hämeenlinna

Fig. 6. The Contextual Reader interface targeting the Kansa Taisteli magazine articles

The purpose of the perspective is two-fold: 1) to help a user find 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. 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, 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 article appears on the screen in the CORE Contextual Reader interface [14]. Depicted in Fig. 6, 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, data is shown to the user, providing contextual information for an enhanced reading experience. In Fig. 6, for example, detailed data are shown about *Raymond August Ericsson*, one of the battalion commanders discussed in the article.

Solving the technical issues, however still left the problem of semantic disambiguation; in this case this concerned named entity recognition of correct people and military units. The identification was made by customizing the SPARQL queries, the order of the queries, and the article metadata. Each magazine article was identified and firstly references to people were searched from the text. The identification of people was done by using name and possibly a rank. Secondly the linking of the military units was performed from the remaining text. The article metadata was also used to identify the war to which the events of the article are related to. Afterwards the military units were linked based on the war into the corresponding units. A detailed description and evaluation of the process is available at [17].

5.4 Photographs

WarSampo contains a dataset of the metadata of ca. 160 000 historical photographs taken by Finnish soldiers during WWII. The data contains e.g. captions of the photographs. The actor ontology was used to automatically disambiguate and link people and military units mentioned in the metadata. Information in the actor ontology was used extensively in linking: For example, when disambiguating people, names, ranks, promotion dates, military units, sources, medals, and death dates were used to rank otherwise ambiguous mentions in the photograph captions. [7]

The results of the linking can be seen in the person and unit perspectives of the WarSampo portal, as well as in the photograph perspective itself[35] which provides a faceted search interface for the photographs.

6 Related work, and Discussion

There are several projects publishing linked data about the World War I on the web, such as Europeana Collections 1914–1918[36], 1914–1918 Online[37], WW1 Discovery[38], Out of the Trenches[39], Muninn [19], and WW1LOD [15]. There are few works that use the Linked Data approach to World War II, such as [3,1], Defence of Britain[40], and Open Memory Project[41]. The main focus on our work is on representing an actor as a biographical life story, unlike databases like Getty ULAN or Smithsonian American Art Museum [16] that have actor vocabularies.

Our research group, Semantic Computing Research Group (SeCo), has produced several projects with highly interlinked actor ontologies: The National Biography, CultureSampo[42], BookSampo[43], and Norssit—High School Alumni [11] datasets. Bio CRM model[44] is developed to facilitate and harmonize the representation of an actor in semantic web, and therefore deals with the same problematics as the WarSampo actor ontology.

We have considered combining the different datasets like articles and photographs to actor ontology as one of the use-cases of the actor ontology. The evaluation of the

[35] http://www.sotasampo.fi/en/photographs

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

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

[38] http://ww1.discovery.ac.uk

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

[40] http://thesaurus.historicengland.org.uk/thesaurus.asp?thes_no=365&thes_name=Defence%20of%20Britain%20Thesaurus

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

[42] http://seco.cs.aalto.fi/applications/kulttuurisampo/

[43] http://seco.cs.aalto.fi/applications/kirjasampo/

[44] http://seco.cs.aalto.fi/projects/biographies/

ontology and actor dataset, has been work- and data-driven e.g. it has developed to the needs of semantically representing the data and of rendering the data at the end-user portal. 94 percent of users come from Finland and 25 percent of them are returning visitors. We have received feedback via the user interface, and we have considered their comments e.g. on misidentified people.

Main requirement for the ontology was to represent changes in spatio-temporal context as described in Introduction. Constant actor resources are enriched with events marking the changes in spatio-temporal continuity, adding details to the semantic biographical representation, and connecting the otherwise separate datasets of personnel, units, places, articles, photographs etc. The unit model had to be capable of representing even more dynamical changes than with people; identifiers like name and abbreviation may change in the time domain. The army hierarchy is represented as a tree graph where the groups are connected by the events of joining.

The actor ontology is based on CIDOC CRM standard which provides a clear framework and basis for actor-event schema. The Muninn Military Ontology offered an example of modeling military concepts semantically. In conclusion, there was no obvious basis for the ontology. On the contrary, it was constructed by combining principles of several solutions all serving different needs.

In a similar way Warsampo project has collected historical, wartime information from Finland. There is abundance of information about the WW2 in different countries, written in local languages, and published in various formats; often even having divergent points of view. Collecting the data and publishing it as LOD forms a tremendous field of work, but aims at constructing a comprehensive, worldwide database. In the events of history, individual people and groups are at the focal center; it is from their point of view that we build our notion of history.

The ontology model represented in this article may not be all-purpose suitable, but we encourage and hope to inspire the researchers to develop the ideas further.

Acknowledgements Our work is funded by the Open Science and Research Initiative[45] of the Finnish Ministry of Education and Culture, the Finnish Cultural Foundation, and the Academy of Finland.

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. Bunescu, R.C., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: EACL. vol. 6, pp. 9–16 (2006)
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–Verlag (November 2005)
4. Doerr, M.: The CIDOC CRM – an ontological approach to semantic interoperability of metadata. AI Magazine 24(3), 75–92 (2003)

[45] http://openscience.fi/

5. Guha, R.V., Brickley, D., Macbeth, S.: Schema. org: Evolution of structured data on the web. Communications of the ACM 59(2), 44–51 (2016)
6. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R.: Evaluating entity linking with Wikipedia. Artificial Intelligence 194, 130–150 (Jan 2013), http://dx.doi.org/10.1016/j.artint.2012.04.005
7. 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 2017. June 19-20, Galway, Ireland. Springer-Verlag (2017), in press
8. 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
9. Hyvönen, E.: Cultural heritage linked data on the semantic web: Three case studies using the sampo model (2017), http://seco.cs.aalto.fi/publications/submitted/hyvonen-vitoria-2017.pdf, invited talk, Proceedings of the VIII Encounter of Documentation Centres of Contemporary Art: Open Linked Data and Integral Management of Information in Cultural Centres Artium, Vitoria-Gasteiz, Spain, 2016. Forthcoming
10. Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J., Mäkelä, E.: WarSampo Data Service and Semantic Portal for Publishing Linked Open Data about the Second World War History. In: The Semantic Web – Latest Advances and New Domains (ESWC 2016). pp. 758–773. Springer-Verlag (2016)
11. Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., Sirola, L.: Reassembling and enriching the life stories in printed biographical registers: Norssi high school alumni on the semantic web. In: Proceedings, Language, Technology and Knowledge 2017. June 19-20, Galway, Ireland. Springer-Verlag (February 2017), http://ldk2017.org/, accepted
12. 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: Sack, H., Rizzo, G., Steinmetz, N., Mladenić, D., Auer, S., Lange, C. (eds.) The Semantic Web: ESWC 2016 Satellite Events. Springer-Verlag (June 2016)
13. Mäkelä, E.: Combining a rest lexical analysis web service with sparql for mashup semantic annotation from text. In: European Semantic Web Conference. pp. 424–428. Springer (2014)
14. Mäkelä, E., Lindquist, T., Hyvönen, E.: CORE - a contextual reader based on linked data. In: Proceedings of Digital Humanities 2016, long papers (July 2016)
15. 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.
16. Szekely, P., Knoblock, C.A., Yang, F., Zhu, X., Fink, E.E., Allen, R., Goodlander, G.: Connecting the Smithsonian American Art Museum to the Linked Data Cloud. In: Extended Semantic Web Conference. pp. 593–607. Springer (2013)
17. 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: Proceedings, Language, Technology and Knowledge 2017. June 19-20, Galway, Ireland. Springer-Verlag (February 2017), accepted
18. Taylor, A.: Introduction to cataloging and classification. Library and Information Science Text Series, Libraries Unlimited (2006)
19. Warren, R.: Creating specialized ontologies using Wikipedia: The Muninn experience. Berlin, DE: Proceedings of Wikipedia Academy: Research and Free Knowledge (WPAC2012). URL: http://hangingtogether. org (2012)

Publication III

Mikko Koho, Lia Gasbarra, Jouni Tuominen, Heikki Rantala, Ilkka Jokipii, and Eero Hyvönen. AMMO Ontology of Finnish Historical Occupations. In *Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH 2019), Rome, Italy, June 3, 2019*, Antonella Poggi (editor), CEUR Workshop Proceedings, volume 2375, pages 91–96, ISSN 16130073, online CEUR-WS.org/Vol-2375/short2.pdf, June 2019.

AMMO ontology of Finnish historical occupations

Mikko Koho[1] mikko.koho@aalto.fi
Lia Gasbarra[1] lia.gasbarra@aalto.fi
Jouni Tuominen[2,1] jouni.tuominen@helsinki.fi
Heikki Rantala[1] heikki.rantala@aalto.fi
Ilkka Jokipii[3,4] ilkka.jokipii@arkisto.fi
Eero Hyvönen[2,1] eero.hyvonen@helsinki.fi

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

Abstract

This paper introduces AMMO Ontology of Finnish Historical Occupations. AMMO is based on thousands of occupational labels extracted from three Finnish military historical datasets of the early 20th century: the first consists of the ca. 40 000 war-related death records around the time of the Finnish Civil War (1914–1922); the second consists of the ca. 95 000 death records of Finnish soldiers in the Second World War (1939–1945); the third contains the ca. 4500 records of Finnish prisoners of war in the Soviet Union during the WW2. Our goal from a Digital Humanities perspective is to use AMMO to study military history and these datasets based on the occupation and social status of the soldiers. AMMO will also be used as a component for faceted search and semantic recommendation in two semantic portals for Finnish military history. AMMO is aligned with the international historical occupation classification HISCO and with a modern Finnish occupational classification for international and national interoperability. The ontology is published as Linked Open Data in an ontology service.

1 Introduction

The measurement of historical social stratification has been a source of discussion in social history studies in the last decades [14]. Whether skills, capital, property, nobility, or occupational prestige is a suitable measure of social status has been debated: often researchers work with vague occupational information, as occupational labels are unclear, and there might not be enough context information to understand the reality of the people working in the occupation.

After extensive research on large historical datasets, the HISCO historical international standard classification of occupations [19] was published in 2002. It provides an international comparative classification system of history of work, particularly for occupational titles in the 19th and early 20th centuries. HISCO encodes not only occupation, but also information about prestige, property and family relations can be included. In general, the national classifications of occupations or census tables, differ in structure and detail within a country, and especially in international context. HISCO provides a tool for transnational comparative studies while also enabling the harmonization of occupations in censuses and datasets on a national scale.

In: Carlo Meghini, Antonella Poggi (eds.): Proceedings of ODOCH 2019 – Open Data and Ontologies for Cultural Heritage – Rome, Italy, 3 June 2019, published at http://ceur-ws.org

AMMO ontology will provide a harmonized view of Finnish historical occupations, which is linked to HISCO classification. The AMMO background and involved manual expert work has been discussed in a previous publication [2]. This paper builds upon the previous work to present the processes used to create the ontology, the ontology design rationale, and the ontology model. HISCO provides the hierarchical backbone of occupational groups in AMMO, as well as social stratification information through several measures like HISCLASS [18,11], a HISCO-based 12 level social classification system, and HISCAM [10,11], a social interaction distance measure. AMMO is also aligned with the Finnish Classification of occupations 1980 [17] (COO1980), a social stratification classification system in use in Finland.

AMMO ontology is based on occupational labels extracted from three Finnish military historical datasets of the early 20th century: the first consists of the ca. 40 000 war-related deaths around the time of the Finnish Civil War (1914–1922)[1]; the second consists of the ca. 95 000 death records of Finnish soldiers in the Winter War and Continuation War (1939–1945) [8]; the third contains the ca. 4500 records of the Finnish prisoners of war in the Soviet Union during the WW2. The two latter are part of the WarSampo[2] data service and semantic portal [4].

Motivation for AMMO comes from two separate usage scenarios. First is using the occupations in a user interface with a faceted search and the second is performing historical research on datasets consisting of data about people. Using the raw occupational labels does not enable the selection of person records based e.g. on the occupational field, social status, and various spellings of a single occupation. These issues can be solved by organizing the occupational labels into an ontology and linking to classifications with information on the social status of the occupation.

The benefits of an occupation ontology in the two scenarios can be summarized as follows:

- **User-interfaces.** User-interfaces employing faceted search [13] (e.g. semantic portals) benefit from organizing each facets' selection into a controlled vocabulary. This holds also for other user interface designs that list or show all of the values within a dataset to a user. Occupations are one of the key variables in many fields of history [19], and thus are one of the natural facets when exploring, studying and analyzing a dataset consisting of people. Using an ontology of occupations enables showing and using hierarchical facet options, and to group synonyms together into a single option and separate homonyms into separate options. Combined with information on the social stratification related to each occupation, we are able to create additional facets based on the social classes.
- **Historical research.** Digital Humanities researchers studying and researching history can use the ontology to get more understanding about the social stratification and occupational distribution within a dataset. Combined with ontology-based query expansion [22], the ontology enables the selection and comparative study of people and their information, based on arbitrary grouping resources, like the occupational field and social class. Also, these prosopographical groups [20] can be enriched with information like the average social class, and the most common occupational field. Many of the research questions of a collaborating historian revolve around social stratification, which is feasible to study only after linking the occupational labels to social stratification measures or classes. An example of the research questions we are trying to answer is ”what is the difference in the social stratification of the two sides fighting in the Finnish Civil War? Which social strata have joined either side in the war in different parts of the country?”

Our work is based on earlier studies about classifying occupations and social stratification. We strive to use pre-existing classifications as much as possible, so surveying the existing occupation classifications has been fruitful, and it sets the limits of the work, as manual expert work on vocabularies is time-consuming.

2 Existing Classifications of Historical Occupations

HISCO is based on a pre-existing international classification of occupations: ISCO-68, which in many countries has been adopted as a guideline for the creation of a national occupation classification scheme. In that case, the aligning of a pre-existing national classification scheme into HISCO is less problematic, since the structures are similar and entries are easily comparable. There is a Finnish version of the Nordic classification of occupations from 1963 [9], based on ISCO-58, which the ISCO-68 is based on. A newer Finnish classifications of occupations from 1980 [17], used e.g. in late 20th century census data, is in turn based on the aforementioned Nordic classification of occupations.

[1] http://www.ldf.fi/dataset/narc-sotasurmat1914-22

[2] http://www.ldf.fi/dataset/warsa

The HISCO encoding process in AMMO is carried out manually: occupational labels are linked to HISCO using the COO1980 as a reference, which is a consistent source of about 5100 specific occupational terms arranged hierarchically. The detailed occupational entries and description of the occupational groups in it helped to interpret and understand the numerous labels enough to enable the HISCO coding. Some occupations have required more specific attention, e.g. those with uncertain attribution such as *"keittäjä"*, cook, which presents many alternatives like *canteen cook*, *sugar cooker*, *sterilizing cook* or *pulp digester operator*, and distinct but hermetic occupational names, such as *"happomies"*, literally "acid man", which is a specific pulp industry worker.

Interesting sources of data for comparative studies are the national censuses of the early 20th century. These, however, group occupations under large, coarse categories, which are impossible to directly link to AMMO or HISCO, as the actual occupations are not known.

Other Finnish historical sources presenting listings of occupations are, for example, the yearly classifications of worker occupations for bread voucher distribution from 1940 [5] to 1943 [6] where working population was divided by occupation and the production sector. Population was ranked according to the grade of manual labour performed; harder labour corresponded to a higher class of bread, butter, and milk voucher. Specifically, the classification of 1943 [6] presents very detailed listings of occupations, accompanied by the corresponding value of the voucher. It is evident how the purpose of a classification influences its intrinsic structure and level of detail.

3 Creating an Ontology of Finnish Historical Occupations

Table 1. Datasets providing Finnish historical occupations for AMMO.

Name	Data provider	Persons	Occupations
WW1 War Victims	National Archives	39 931	1391
WW2 Death Records	National Archives	94 700	2155
WW2 Prisoners of War	National Prisoners of War Project	4460	576

The source datasets of AMMO are presented in Table 1, containing information of ca. 139 000 historical persons (soldiers), of which almost all are annotated with at least one occupational labels, summing up to thousands of different occupation titles. In the datasets, alternative or abridged forms of the same occupational title are often present (for example: *"hitsari"* and *"hitsaaja"* for welder). In some cases, the occupational label of a person is actually not an occupation but a social role, honorary title, degree or status, such as *student*, *nobleman*, *child*, *tenant* or *master of science*. Many children are labeled under their father's occupation, such as *driver's son*. Although occupations in HISCO are by definition solely activities that generate a remuneration, it is possible to also categorize many social roles or statuses through HISCO relation and status coding.

A common approach to creating an ontology model is to reuse existing non-ontological knowledge resources such as thesauri, classification schemes and lexicons, or ontological knowledge resources [21]. For AMMO, the existing Finnish classification of occupations 1980 [17] was used as both a thesaurus, and a classification scheme for the identification of both a specific occupation and a social status.

The main design rationale of the ontology model comes from the aforementioned two usage scenarios, i.e. the need to use occupational information in faceted search and historical research. We have striven to create the simplest possible model to provide results for these, that does not lose important information given in the occupational labels. The secondary goal is to provide a useful artifact for anyone studying or analyzing historical data containing people with Finnish language occupational labels.

One approach to achieving the needed HISCO-linking would be to annotate the HISCO occupations directly with the corresponding Finnish occupational labels found in our datasets. However, as the HISCO status and relationship variables are an important part of the HISCLASS coding [18], performing the coding on only HISCO occupation code would be erroneous for many occupational labels, as e.g. a pharmacy student would be considered having the same HISCLASS code as a pharmacist.

The AMMO ontology consists of individual SKOS concepts [12], each depicting one occupation with synonyms and alternative spellings gathered to the same concept as alternative labels. This enables to fully employ HISCLASS coding, and to keep the level of detail used in the occupational labels in the source datasets. The AMMO concepts are further separated into 5 classes depending on whether the occupational label refers to 1) an actual occupation, 2) a degree, 3) an honorary title, 4) a military rank, or 5) a social role.

The ontology model is presented in Figure 1 through two example resources, of which one is an occupation (pharmacist), and one is a social role related to the occupation (pharmacy student). RDF resources are depicted as ellipses, literals as rectangles and related datasets as clouds. The figure displays the linkage to the existing occupation classifications, and the related classification hierarchies. There are no direct relations between the AMMO occupation concepts, but the concepts (in green) are linked to other resources:

- HISCO (in blue), which contains the occupation hierarchy, relationship code, status code, HISCAM measure, and HISCLASS class,
- COO1980 (in red), which contains the occupation hierarchy, socioeconomic status class,
- KOKO ontology (in yellow).

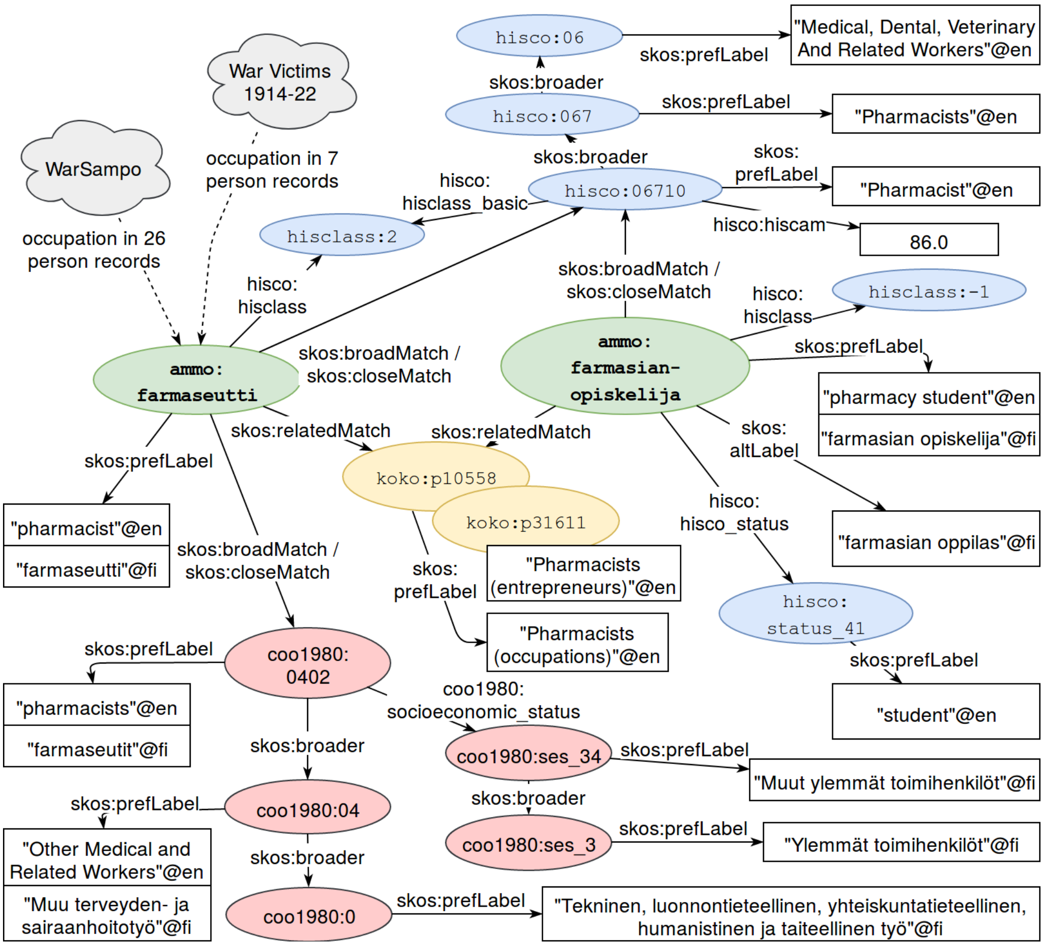


Fig. 1. The AMMO ontology model with two example AMMO resources.

In Figure 1, the namespace prefix *ammo:* refers to AMMO ontology namespace, *hisco:* refers to the RDF conversion of HISCO, *coo1980:* refers to the RDF conversion of COO1980, and *koko:* refers to the KOKO ontology. The property *hisco:hisclass* annotates the HISCLASS class code (1-12, or -1 for no occupation) of an AMMO occupation, whereas the base HISCLASS code of an HISCO occupation is given with the property *hisco:hisclass_basic*. There are two linked KOKO ontology concepts for both AMMO resources.

The overall process of creating the AMMO ontology is as follows:

1. Combining occupational labels from the datasets, and automatic grouping of easily identifiable synonyms,
2. The manual harmonization of the occupational labels and linking to external vocabularies,
3. Transforming the occupations into a SKOS vocabulary,
4. Validating and refining the ontology as needed,
5. Integrating HISCO and COO1980 classifications as linked SKOS vocabularies.

In step 1, the occupational labels are extracted from the datasets, and programmatically harmonized using a few simple rules to group occupational labels containing common interchangeable worker names *”työläinen”*, *"työmies"*, and *"työntekijä"*, which in most cases are used for identical meaning, and occupations with almost identical labels based on a Jaro-Winkler string similarity limit of 0.97. This results in a flat vocabulary of 2053 distinct occupations, containing a total of 2977 distinct occupational labels.

Step 2 begins with transforming the flat vocabulary into a spreadsheet, for an ontology developer to work on. The ontology developer re-engineers the ontology to account for synonymy, while manually linking the occupations to HISCO and COO1980 classifications and to the KOKO ontology[3].

Step 3 consists of RML [1] transformation of the spreadsheet into a SKOS [12] vocabulary. The ontology already contains URI references to the used classifications and the KOKO ontology, as well as annotations of preferred and alternative labels.

In step 4, the created ontology is validated and refined as needed. One key validation is to link the person records in the source datasets to AMMO, and inspect the results.

Step 5 consists of transforming the HISCO version 2018.01 [11] and the COO1980 main hierarchy into SKOS vocabularies, and enriching them with pre-existing English and Finnish labels. They are integrated into AMMO to provide hierarchical backbones, which might still reveal a need to refine the manual harmonization.

4 Discussion

This paper presented the foundations of the AMMO ontology, which will enable better utilization of Finnish historical datasets containing information about people. User interfaces can make use of either of the occupational hierarchies to provide a faceted search of people based on their occupation, in addition to enabling selection based on persons' social class, or e.g. the line of work (agriculture, metal workers, etc.). Historians can pursue research questions related to social stratification, line of work, and various occupational groups.

AMMO is an ontological representation of Finnish occupations, for the period ranging from 1914 to 1945, therefore having clear boundaries in space and time.

In order to link occupational names gathered from disparate sources into HISCO coding, the effort of interpretation and attentive adjustments are necessary, despite historically relevant occupational statistics and official classifications of occupations being readily available. The first half of the twentieth century saw substantial transformations in the Finnish society, especially in the agricultural sector: a long-lived vertical hierarchical system was shifting towards a more horizontal structure. The statuses of some agricultural occupations have changed dramatically while the occupation name has remained the same. This semantic drift [3] in the occupations causes the HISCO codings to be time-dependent, and HISCO coding based on occupations in the early 20th century might not be accurate in previous centuries. In addition to being a possible obstacle to some comparisons, the semantic drift provides an interesting topic to study in the future.

One interesting direction of research would be to compare the social stratification of people on different sides of the Finnish civil war with that of the social stratification on the national level. This would require at least to estimate a HISCLASS level to each coarse-grained occupational group.

Generally, Finland presents an ideal situation in population data availability and accuracy [7]. The first population census was completed already in 1749 under Swedish jurisdiction, after which they have been regularly redone [16]. Population registrations have been historically also registered in detail [15].

Currently work on AMMO is in step 3 of the process depicted in Section 3. Later, the AMMO ontology, along with the conversion pipeline will be published online for anyone to use.

References

1. Dimou, A., Sande, M.V., Slepicka, J., Szekely, P., Mannens, E., Knoblock, C., Walle, R.V.D.: Mapping hierarchical sources into RDF using the RML mapping language. Proceedings - 2014 IEEE International Conference on Semantic Computing, ICSC 2014 pp. 151–158 (2014). https://doi.org/10.1109/ICSC.2014.25
2. Gasbarra, L., Koho, M., Jokipii, I., Rantala, H., Hyvönen, E.: An ontology of finnish historical occupations. In: Proceedings of the 16th ESWC Conference (ESWC 2019), Posters & demonstrations. Springer (June 2019)
3. Gulla, J.A., Solskinnsbakk, G., Myrseth, P., Haderlein, V., Cerrato, O.: Semantic drift in ontologies. In: WEBIST (2). pp. 13–20 (2010)
4. Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J., Mäkelä, E.: WarSampo Data Service and Semantic Portal for Publishing Linked Open Data about the Second World War History. In: The Semantic Web – Latest Advances and New Domains (ESWC 2016). pp. 758–773. Springer-Verlag (2016)
5. Kansanhuoltoministeriö: Ohjeet leipä-, rasva- ja maitokorttien jakelusta. Kansanhuoltoministeriö, Helsinki (1940)

[3] http://finto.fi/koko/en/

6. Kansanhuoltoministeriö: Leipäkorttien jaossa noudatettava ammattiryhmittely. Kansanhuoltoministeriö, Helsinki (1943)
7. Kinnunen, M.: Luokiteltu sukupuoli. Vastapaino, Tampere (2001)
8. 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 2017 Satellite Events. pp. 369–383. Springer (2017)
9. Kulkulaitosten ja yleisten töiden ministeriön työvoima-asiain osasto: Pohjoismainen ammattiluokittelu, suomenkielinen laitos. Valtioneuvoston kirjapaino, Helsinki (1963)
10. Lambert, P., Zijdeman, R., Leeuwen, M.V., Maas, I., Prandy, K.: The construction of HISCAM: A stratification scale based on social interactions for historical comparative research. Historical Methods: A Journal of Quantitative and Interdisciplinary History **46**(2), 77–89 (2013)
11. Mandemakers, K., Mourits, R.J., Muurling, S., Boter, C., van Dijk, I.K., Maas, I., de Putte, B.V., Zijdeman, R.L., Lambert, P., van Leeuwen, M.H., van Poppel, F., Miles, A.: HSN standardized, HISCO-coded and classified occupational titles, release 2018.01. IISG, Amsterdam (2018)
12. Miles, A., Matthews, B., Wilson, M., Brickley, D.: SKOS core: simple knowledge organisation for the web. In: International Conference on Dublin Core and Metadata Applications. pp. 3–10 (2005)
13. Oren, E., Delbru, R., Decker, S.: Extending Faceted Navigation for RDF Data. In: The Semantic Web - ISWC 2006. pp. 559–572. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
14. Van de Putte, B., Buyst, E.: Occupational titles? hard to eat, easy to catch. Journal of Belgian History (JBH) **40**(1-2), 7–31 (2010)
15. Sköld, P.: The birth of population statistics in sweden. The History of the Family **9**(1), 5–21 (2004)
16. Statistics Finland: Historiallisen tilastotiedon opas. https://guides.stat.fi/historiallisentilastotiedonopas/vaestolaskennat, accessed: 2019-05-14
17. Statistics Finland: Classification of Occupations 1980. Käsikirjoja / Tilastokeskus, Statistics Finland, Helsinki (1981)
18. Van Leeuwen, M.H.D., Maas, I.: HISCLASS: A historical international social class scheme. Leuven University Press (2011)
19. Van Leeuwen, M.H.D., Maas, I., Miles, A.: HISCO: Historical international standard classification of occupations. Leuven University Press (2002)
20. Verboven, K., Carlier, M., Dumolyn, J.: A short manual to the art of prosopography. In: Prosopography approaches and applications. A handbook, pp. 35–70. Unit for Prosopographical Research (Linacre College) (2007)
21. Villazón-Terrazas, B.C., Suárez-Figueroa, M., Gómez-Pérez, A.: A pattern-based method for re-engineering non-ontological resources into ontologies. International Journal on Semantic Web and Information Systems (IJSWIS) **6**(4), 27–63 (2010)
22. Voorhees, E.M.: Query expansion using lexical-semantic relations. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 61–69. Springer-Verlag New York, Inc. (1994)

Publication IV

Eero Hyvönen, Erkki Heino, Petri Leskinen, Esko Ikkala, Mikko Koho, Minna Tamper, Jouni Tuominen, and Eetu 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: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29 – June 2, 2016, Proceedings*, Harald Sack, Eva Blomqvist, Mathieu d'Aquin, Chiara Ghidini, Simone Paolo Ponzetto, and Christoph Lange (editors), Lecture Notes in Computer Science, volume 9678, pages 758–773, ISBN 9783319341286, Springer, Cham, May–June 2016.

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, Eetu Mäkelä

Semantic Computing Research Group (SeCo), Aalto University, Finland
http://seco.cs.aalto.fi/, firstname.lastname@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 [5,3] 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 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.

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

[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.

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

#	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	3,357 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

Table 1. Central datasets of WarSampo.

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.

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.

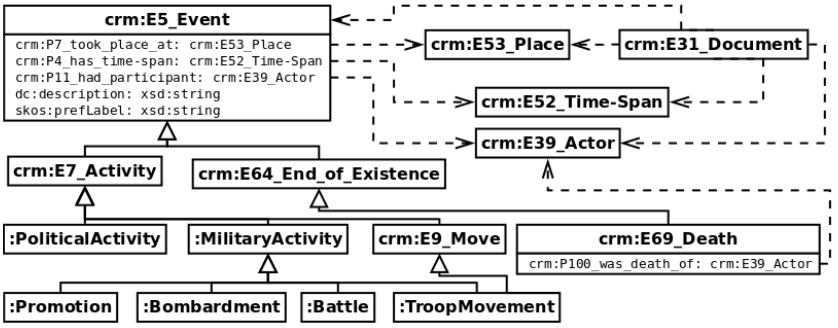


Fig. 1. Core classes of CIDOC CRM used in WarSampo.

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, re-

[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

spectively. 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, 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 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.

[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

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.

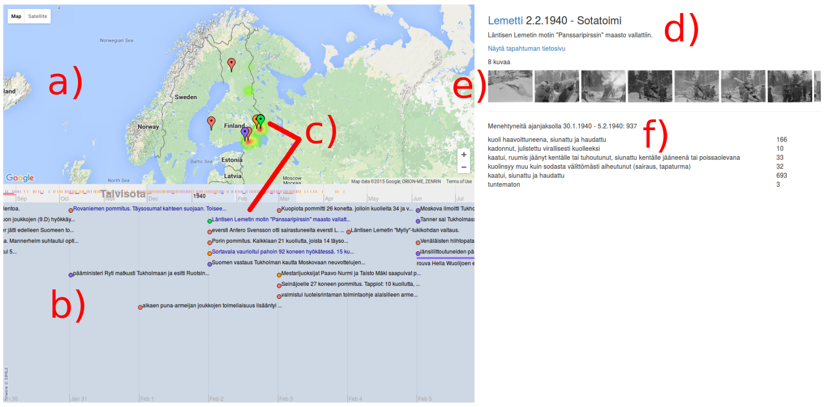
The data service can be used as a basis for Rich Internet Applications (RIA). A demonstration of this is the WarSampo Portal, where *all* functionality is implemented on the client side using JavaScript, only data is fetched from the server side SPARQL endpoints. In below, the six perspectives of the 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 description and contextualizing hyperlinks to other perspectives through entities linked to the event.

Fig. 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

[14] http://ldf.fi/warsa/sparql

[15] http://www.sotasampo.fi/events



lentoa.
Rovaniemen pommitus. Täysosumat kahteen suojaan. Toisee...
Kuopiota pommitti 26 konetta. jolloin kuolleita 34 ja v...
Moskova ilmoitti Tukho
uon joukkojen (9.D) hyökkäy...
Läntisen Lemetin motin "Panssaripirssin" maasto vallatt...
Tanner sai Tukholmas
r jätti edelleen Suomeen to...
eversti Antero Svensson otti sairastuneelta eversti L. ...
Läntisen Lemetin "Mylly"-tukikohdan valtaus.
a. Mannerheim suhtautui opti...
Porin pommitus. Kaikkiaan 21 kuollutta, joista 14 täyso...
Venäläisten hiihtopata
ui 5...
Sortavala vaurioitui pahoin 92 koneen hyökätessä. 15 ku...
länsiliittoutuneiden pä
Suomen vastaus Tukholman kautta Moskovaan neuvottelujen...
rouva Hella Wuolijoen e
pääministeri Ryti matkusti Tukholmaan ja esitti Ruotsin...
Mestarijuoksijat Paavo Nurmi ja Taisto Mäki saapuivat p...
Seinäjoelle 27 koneen pommitus. Tappiot: 10 kuollutta, ...
valmistui luoteisrintaman toimintaohje alaisilleen arme...
alkaen puna-armeijan joukkojen toimeliaisuus lisääntyi ...

Lemetti 2.2.1940 - Sotatoimi

Läntisen Lemetin motin "Panssaripirssin" maasto vallattiin.

Näytä tapahtuman tietosivu

8 kuvaa

Menehtyneitä ajanjaksolla 30.1.1940 - 5.2.1940: 937

kuoli haavoittuneena, siunattu ja haudattu	166
kadonnut, julistettu virallisesti kuolleeksi	10
kaatui, ruumis jäänyt kentälle tai tuhoutunut, siunattu kentälle jääneenä tai poissaolevana	33
kuolinsyy muu kuin sodasta välittömästi aiheutunut (sairaus, tapaturma)	32
kaatui, siunattu ja haudattu	693
tuntematon	3

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

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 [13,12]. 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.

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

[16] http://sotasampo.fi/persons

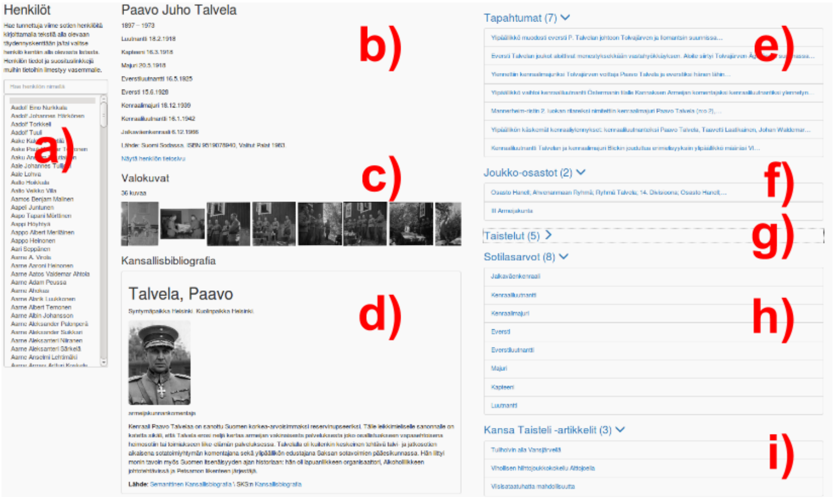


Fig. 3. Person perspective.

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 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

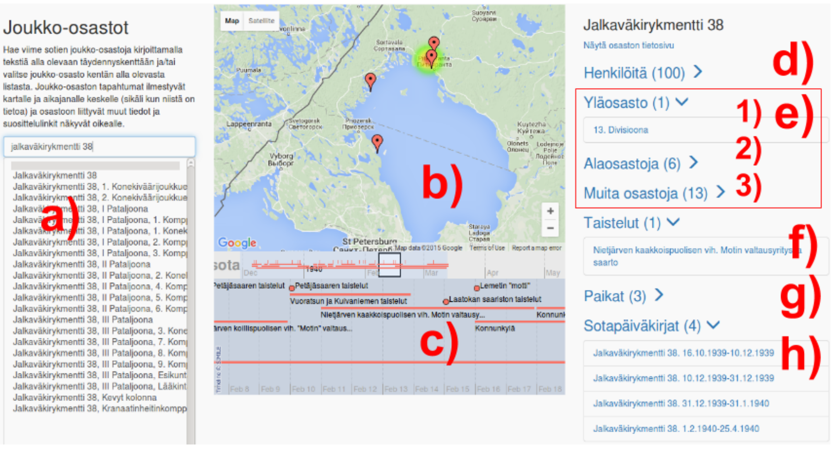
[17] http://sa-kuva.fi/neo?tem=webneoeng

[18] http://www.ldf.fi/dataset/history

(*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*.



Joukko-osastot

Hae viime sotien joukko-osastoja kirjoittamalla tekstiä alla olevaan täydennyskenttään ja/tai valitse joukko-osasto kentän alla olevasta listasta. Joukko-osaston tapahtumat ilmestyvät kartalle ja aikajanalle keskelle (sikäli kun niistä on tietoa) ja osastoon liittyvät muut tiedot ja suosittelulinkit näkyvät oikealle.

jalkaväkirykmentti 38

Jalkaväkirykmentti 38
Jalkaväkirykmentti 38, 1. Konekiväärijoukkue
Jalkaväkirykmentti 38, 2. Konekiväärijoukkue
Jalkaväkirykmentti 38, I Pataljoona
Jalkaväkirykmentti 38, I Pataljoona, 1. Komp
Jalkaväkirykmentti 38, I Pataljoona, 1. Konek
Jalkaväkirykmentti 38, I Pataljoona, 2. Komp
Jalkaväkirykmentti 38, I Pataljoona, 3. Komp
Jalkaväkirykmentti 38, II Pataljoona
Jalkaväkirykmentti 38, II Pataljoona, 2. Konel
Jalkaväkirykmentti 38, II Pataljoona, 4. Komp
Jalkaväkirykmentti 38, II Pataljoona, 5. Komp
Jalkaväkirykmentti 38, II Pataljoona, 6. Komp
Jalkaväkirykmentti 38, III Pataljoona
Jalkaväkirykmentti 38, III Pataljoona, 3. Kone
Jalkaväkirykmentti 38, III Pataljoona, 7. Kom
Jalkaväkirykmentti 38, III Pataljoona, 8. Kom
Jalkaväkirykmentti 38, III Pataljoona, 9. Kom
Jalkaväkirykmentti 38, III Pataljoona, Esikunt
Jalkaväkirykmentti 38, III Pataljoona, Lääkint
Jalkaväkirykmentti 38, Kevyt kolonna
Jalkaväkirykmentti 38, Kranaatinheitinkompp

Jalkaväkirykmentti 38

Näytä osaston tietosivu

Henkilöitä (100) >

Yläosasto (1) ⌄

13. Divisioona

Alaosastoja (6) >

Muita osastoja (13) >

Taistelut (1) ⌄

Nietjärven kaakkoispuolisen vih. Motin valtausyritys ja saarto

Paikat (3) >

Sotapäiväkirjat (4) ⌄

Jalkaväkirykmentti 38. 16.10.1939-10.12.1939
Jalkaväkirykmentti 38. 10.12.1939-31.12.1939
Jalkaväkirykmentti 38. 31.12.1939-31.1.1940
Jalkaväkirykmentti 38. 1.2.1940-25.4.1940

Fig. 4. Army unit perspective.

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. The results

[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

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.

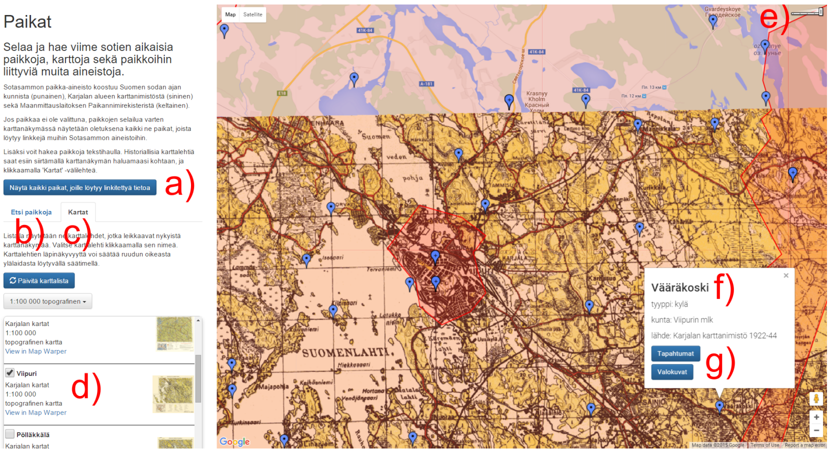
The data consists of over 3,000 Finnish army units, including Land Forces, Air Forces, Navy and its vessels, Medical Corps, stations of Anti-Aircraft Warfare and Skywatch, Finnish White Guard, and Swedish Volunteer Corps. The main sources of information have been the War Diaries and Organization Cards. 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.

Fig. 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

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



Paikat

Selaa ja hae viime sotien aikaisia paikkoja, karttoja sekä paikkoihin liittyviä muita aineistoja.

Sotasammon paikka-aineisto koostuu Suomen sodan ajan kunnista (punainen), Karjalan alueen karttanimistöstä (sininen) sekä Maanmittauslaitoksen Paikannimirekisteristä (keltainen).

Jos paikkaa ei ole valittuna, paikkojen selailua varten karttanäkymässä näytetään oletuksena kaikki ne paikat, joista löytyy linkkejä muihin Sotasammon aineistoihin.

Lisäksi voit hakea paikkoja tekstihaulla. Historiallisia karttalehtiä saat esiin siirtämällä karttanäkymän haluamaasi kohtaan, ja klikkaamalla 'Kartat' -välilehteä.

Listalla näytetään ne karttalehdet, jotka leikkaavat nykyistä karttanäkymää. Valitse karttalehti klikkaamalla sen nimeä. Karttalehtien läpinäkyvyyttä voi säätää ruudun oikeasta ylälaidasta löytyvällä säätimellä.

1:100 000 topografinen

Karjalan kartat
1:100 000
topografinen kartta
View in Map Warper

☑ Viipuri
Karjalan kartat
1:100 000
topografinen kartta
View in Map Warper

□ Pölläkkälä
Karjalan kartat

Vääräkoski
tyyppi: kylä
kunta: Viipurin mlk
lähde: Karjalan karttanimistö 1922-44

Fig. 5. Historical places perspective.

”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.

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.

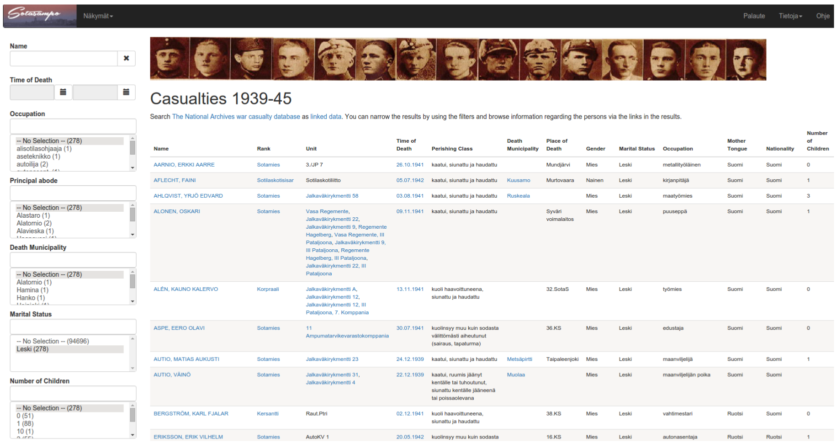
The historical place ontology was created using four data sources: 1) a map application the National Archives of Finland (612 wartime municipalities), 2) Finnish Spatio-Temporal Ontology (polygon boundaries of the municipalities)[24], 3) a dataset of geocoded Karelian map names (35,000 map names with 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.

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

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.



Name

Time of Death

Occupation

-- No Selection -- (278)
alisotilasohjaaja (1)
aseteknikko (1)
autoilija (2)

Principal abode

-- No Selection -- (278)
Alastaro (1)
Alatornio (2)
Alavieska (1)

Death Municipality

-- No Selection -- (278)
Alatornio (1)
Hamina (1)
Hanko (1)

Marital Status

-- No Selection -- (94696)
Leski (278)

Number of Children

-- No Selection -- (278)
0 (51)
1 (88)
10 (1)

Casualties 1939-45

Search The National Archives war casualty database as linked data. You can narrow the results by using the filters and browse information regarding the persons via the links in the results.

Name	Rank	Unit	Time of Death	Perishing Class	Death Municipality	Place of Death	Gender	Marital Status	Occupation	Mother Tongue	Nationality	Number of Children
AARNIO, ERKKI AARRE	Sotamies	3./JP 7	26.10.1941	kaatui, siunattu ja haudattu		Mundjärvi	Mies	Leski	metallityöläinen	Suomi	Suomi	0
AFLECHT, FAINI	Sotilaskotisisar	Sotilaskotiliitto	05.07.1942	kaatui, siunattu ja haudattu	Kuusamo	Murtovaara	Nainen	Leski	kirjanpitäjä	Suomi	Suomi	1
AHLQVIST, YRJÖ EDVARD	Sotamies	Jalkaväkirykmentti 58	03.08.1941	kaatui, siunattu ja haudattu	Ruskeala		Mies	Leski	maatyömies	Suomi	Suomi	3
ALONEN, OSKARI	Sotamies	Vasa Regemente, Jalkaväkirykmentti 22, Jalkaväkirykmentti 9, Regemente Hagelberg, Vasa Regemente, III Pataljoona, Jalkaväkirykmentti 9, III Pataljoona, Regemente Hagelberg, III Pataljoona, Jalkaväkirykmentti 22, III Pataljoona	09.11.1941	kaatui, siunattu ja haudattu		Syväri voimalaitos	Mies	Leski	puuseppä	Suomi	Suomi	1
ALÉN, KAUNO KALERVO	Korpraali	Jalkaväkirykmentti A, Jalkaväkirykmentti 12, Jalkaväkirykmentti 12, III Pataljoona, 7. Komppania	13.11.1941	kuoli haavoittuneena, siunattu ja haudattu		32.SotaS	Mies	Leski	työmies	Suomi	Suomi	0
ASPE, EERO OLAVI	Sotamies	11 Ampumatarvikevarastokomppania	30.07.1941	kuolinsyy muu kuin sodasta välittömästi aiheutunut (sairaus, tapaturma)		36.KS	Mies	Leski	edustaja	Suomi	Suomi	0
AUTIO, MATIAS AUKUSTI	Sotamies	Jalkaväkirykmentti 23	24.12.1939	kaatui, siunattu ja haudattu	Metsäpirtti	Taipaleenjoki	Mies	Leski	maanviljelijä	Suomi	Suomi	1
AUTIO, VÄINÖ	Sotamies	Jalkaväkirykmentti 31, Jalkaväkirykmentti 4	22.12.1939	kaatui, ruumis jäänyt kentälle tai tuhoutunut, siunattu kentälle jääneenä tai poissaolevana	Muolaa		Mies	Leski	maanviljelijän poika	Suomi	Suomi	
BERGSTRÖM, KARL FJALAR	Kersantti	Raut.Ptri	02.12.1941	kuoli haavoittuneena, siunattu ja haudattu		38.KS	Mies	Leski	vahtimestari	Ruotsi	Suomi	0
ERIKSSON, ERIK VILHELM	Sotamies	AutoKV 1	20.05.1942	kuolinsyy muu kuin sodasta		16.KS	Mies	Leski	autonasentaja	Ruotsi	Ruotsi	1

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.

[25] http://www.sotasampo.fi/casualties

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.

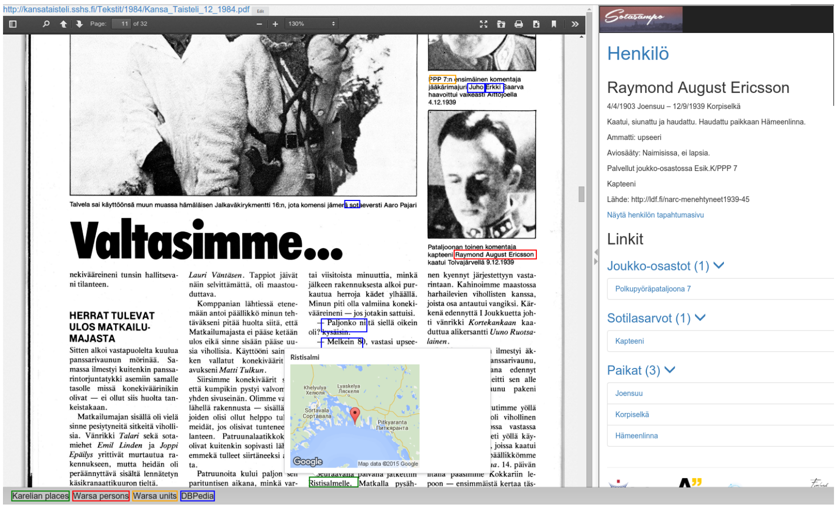
The casualty records were modeled using the class crm:E31_Document with a distinct property for each facet. The property values are annotation resources selected from the corresponding ontologies, such as places. Record instances refer also to events, e.g., the death events of persons.

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*, 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.

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


http://kansataisteli.sshs.fi/Tekstit/1984/Kansa_Taisteli_12_1984.pdf
Page: 11 of 32
130%
Ristisalmi
Map data ©2015 Google

Talvela sai käyttöönsä muun muassa hämäläisen Jalkaväkirykmentti 16:n, jota komensi jämerä sotaeversti Aaro Pajari

Valtasimme...

HERRAT TULEVAT ULOS MATKAILUMAJASTA

PPP 7:n ensimäinen komentaja jääkärimajuri Juho Erkki Saarva haavoittui vaikeasti Aittojoella 4.12.1939

Pataljoonan toinen komentaja kapteeni Raymond August Ericsson kaatui Tolvajärvellä 9.12.1939

Karelian places | Warsa persons | Warsa units | DBPedia

Henkilö

Raymond August Ericsson

4/4/1903 Joensuu – 12/9/1939 Korpiselkä

Kaatui, siunattu ja haudattu. Haudattu paikkaan Hämeenlinna.

Ammatti: upseeri

Aviosääty: Naimisissa, ei lapsia.

Palvellut joukko-osastossa Esik.K/PPP 7

Kapteeni

Lähde: http://ldf.fi/narc-menehtyneet1939-45

Näytä henkilön tapahtumasivu

Linkit

Joukko-osastot (1)

- Polkupyöräpataljoona 7

Sotilasarvot (1)

- Kapteeni

Paikat (3)

- Joensuu
- Korpiselkä
- Hämeenlinna

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

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 spe-

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

cialized 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 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 [2,1] 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

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

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

[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

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: Proc. 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: Proc. of the 4th International Semantic Web Conference (ISWC 2005). Springer–Verlag (November 2005)
3. Crymble, A., Gibbs, F., Hegel, A., McDaniel, C., Milligan, I., Posner, M., Turkel, W.J. (eds.): The Programming Historian. 2 edn. (2015), http://programminghistorian.org/
4. Doerr, M.: The CIDOC CRM – an ontological approach to semantic interoperability of metadata. AI Magazine 24(3), 75–92 (2003)
5. Graham, S., Milligan, I., Weingart, S.: Exploring big historical data. The historian's macroscope. Imperial College Press (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, Morgan & Claypool, 1 edn. (2011), http://linkeddatabook.com/editions/1.0/
7. Hyvönen, E.: Preventing interoperability problems instead of solving them. Semantic Web Journal 1(1–2), 33–37 (December 2010)
8. Hyvönen, E.: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Morgan & Claypool, Palo Alto, CA, USA (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: Proc. of CIDOC 2012 – Enriching Cultural Heritage (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: The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers. pp. 226–230. Springer–Verlag (May 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 Demos. Springer–Verlag (October 2015)
12. Leskinen, J., Juutilainen, A. (eds.): Jatkosodan pikkujättiläinen. WSOY, Finland (2005)
13. Leskinen, J., Juutilainen, A. (eds.): Talvisodan pikkujättiläinen. WSOY, Finland, 4 edn. (2006)
14. Mäkelä, E., Törnroos, J., Lindquist, T., Hyvönen, E.: World War 1 as Linked Open Data (2015), http://seco.cs.aalto.fi/publications/, submitted for review
15. Mäkelä, E.: Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text. In: Proceedings of the ESWC 2014 demonstration track. Springer–Verlag (May 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, Synthesis lectures on information concepts, retrieval, and services, vol. 1. Morgan & Claypool Publishers (2009)

Publication V

Esko Ikkala, Mikko Koho, Erkki Heino, Petri Leskinen, Eero Hyvönen, and Tomi 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), Vienna, Austria, October 22, 2017*, Alessandro Adamou, Enrico Daga, and Leif Isaksen (editors), CEUR Workshop Proceedings, volume 2014, pages 45–56, ISSN 16130073, online CEUR-WS.org/Vol-2014/paper-06.pdf, October 2017.

Prosopographical Views to Finnish WW2 Casualties Through Cemeteries and Linked Open Data

Esko Ikkala[1], Mikko Koho[1], Erkki Heino[1,2], Petri Leskinen[1], Eero Hyvönen[1,2], and Tomi Ahoranta[3]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland and
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
[3] The National Archives of Finland
`http://seco.cs.aalto.fi, http://heldig.fi,`
`http://www.arkisto.fi/en/frontpage`

Abstract. This paper presents an application for studying the death records of WW2 casualties from a prosopograhical perspective, provided by the various local military cemeteries where the dead were buried. The idea is to provide the end user with a global visual map view on the places in which the casualties were buried as well as with a local historical perspective on what happened to the casualties that lay within a particular cemetery of a village or town. Plenty of data exists about the Second World War (WW2), but the data is typically archived in unconnected, isolated silos in different organizations. This makes it difficult to track down, visualize, and study information that is contained within multiple distinct datasets. In our work, this problem is solved using aggregated Linked Open Data provided by the WarSampo Data Service and SPARQL endpoint.

1 Introduction

This paper builds upon the WarSampo project[4] that collects data related to the Finnish WW2, and publishes this data as Linked Open Data [2] on an open SPARQL endpoint. WarSampo is to our best knowledge the first large scale system for serving and publishing WW2 LOD on the Semantic Web.[5]

Seven different user-friendly application perspectives based on the the WarSampo Data Service[6] have been created within the semantic WarSampo Portal[7] [7]. This paper presents a new, 8th application perspective based on the WarSampo Data Service: Cemeteries[8]. This work is based on research done in

[4] `http://seco.cs.aalto.fi/projects/sotasampo/en/`

[5] WarSampo was awarded with the LODLAM Challenge Open Data Prize in Venice, 2017.

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

[7] `http://sotasampo.fi/en`

[8] Premilinary version at `http://www.sotasampo.fi/en/cemeteries`

transforming Finnish WW2 casualties to Linked Data [10] and linking them to other datasets in WarSampo.

Finnish soldiers who perished in WW2 were transported back to their hometown for burial whenever it was possible [11]. Thus, the local cemetery is a natural starting point for studying the common characteristics and events of the residents of one's hometown in the turmoil of the war.

This paper presents new interactive views to the casualty data based on cemeteries and other community level groups. First, all military cemeteries in Finland can be browsed with a faceted browser, and visualized on maps to provide a global view of the cemeteries. Second, another new view is the cemetery information page that provides a community-level view presenting various kinds of information and visualizations about each cemetery and the soldiers buried there, including over 3000 photographs of the cemeteries. Third, the new functionalities of the existing casualties perspective make it possible to study various community level groups. For example, the people buried in a particular cemetery or born in a certain town form a prosopographical [5,13] group which is of interest to local historians. Visualizations of the casualty data can be created providing the end user with insights on what actually happened to these people originating from the same area. Furthermore, the group can be filtered further down into subsets of interest in versatile ways using the faceted browser and new visualizations of the existing Casualties perspective.

The principal use case of the new views is to provide local level views to the casualty data. This use case originates from the data owners at the National Archives of Finland, who emphasized the idea that the users need a local starting point for browsing the data. The existing person perspective of the WarSampo portal provides answers to the question: "What happened to my relatives in WW2" whereas the new views are meant to expose what happened to the user's or the user's relatives' hometown soldiers in WW2.

2 User Groups of WW2 Casualty Data

The data users can roughly be divided into three groups: academic researches, military history enthusiasts and private citizens. First of these groups have widest range of needs regarding the data. On the other hand, they often have the best skills to handle and refine the data by themselves. The focus of academic research regarding this data seems to be shifting from a macro level towards individual and social aspects of war.

Military history enthusiasts usually approach the data from a military unit perspective, or they may concentrate on a certain location during a narrow time frame. They may also be searching for irregularities, such as peaks in numbers of casualties or in certain age groups within the data. One special group of data users in this segment are voluntary researchers, whose aim is to search, locate, and bring back remains of Finnish soldiers who went missing during the battles in the area that is currently part of Russia. They use the data to identify the war victims found during their excavations at the WW2 battle sites.

Private citizens usually begin their search for information with their own relatives who were lost during the war. After finding that out they may go on searching for similar destinies based on age group, unit, or locations (e.g. home towns or the location where their relatives lost their lives).

As mentioned above, academic researchers usually have the best skills to refine the data for their needs. At the other end are private citizens who are usually most dependent on easy-to-manage user interfaces. There have been no studies done on the data user profiles, but it seems apparent that they form the largest group. In this group are also school children who may benefit from this data during their studies as local history aspects are more and more valued in Finnish national curriculum for basic education.

The WarSampo project is targeted at all three user groups. The WarSampo Data Service publishes all information as Linked Open Data so that academic researchers can, at their choice, download the data and process it further, or query it directly with SPARQL. The WarSampo Portal provides user-friendly applications for all the user groups to search, browse, analyze and visualize the data.

3 Cemeteries as Linked Open Data

An important part in studying histories of communities are cemetery studies, which provide a physical record of a community's former inhabitants [12]. In this research we study cemeteries of war graves as data by linking cemeteries to the soldiers who have been buried there, and developing on-line tools and visualizations for analyzing the data.

3.1 Collecting Cemetery Data

A complete listing of all war cemeteries in Finland has not previously been available, but it has been estimated that there are about 690 of them. The Central Organization of Finnish Camera Clubs (Suomen Kameraseurojen Liitto ry, SKsL) is coordinating a project called "War Cemeteries in Finland" where local camera clubs photograph and collect data about all the war cemeteries.

The members of the camera clubs were instructed to take five specific photographs from each cemetery. At the same time the following information will be collected: 1) official name of the cemetery, 2) year of foundation, 3) architect, 4) memorial (name, sculptor and unveiling date), 5) street address, 6) coordinates, 7) former municipality, and 8) current municipality. The cemetery and photograph data is first organized into a CSV table, which is then converted into RDF.[9]

[9] At the time of writing the information collection is still in progress and preliminary data and photographs related to 218 cemeteries have been converted into RDF and published on the WarSampo Data Service.

3.2 Data Model and Linking War Casualties to Cemeteries

For interlinking the heterogeneous datasets of WarSampo, a wartime place ontology containing place names with different levels of granularity and types was created earlier as a pilot implementation of the "Finnish Ontology Service of Historical Places and Maps" [6]. Now, when the cemetery data is added to the place ontology, it brings out a new local level to the ontology, as there is often more than one cemetery within a municipality.

In order to add the cemetery data to the WarSampo Data Service, the event-based data model of WarSampo was extended with a new cemetery class. The part of the WarSampo data model that is relevant to this paper is presented in figure 1. Because the WarSampo data model is founded on the CIDOC Conceptual Reference Model (CRM) [4], the cemetery class is a subclass of the CRM class `E27_Site`. The cemetery data is linked to the person instances via death records and to place ontology via municipalities.

The namespaces `:`, `:crm`, and `:narc` used in figure 1 refer to WarSampo schema (`http://ldf.fi/schema/warsa`), CRM, and Casualties schema (`http://ldf.fi/schema/narc-menehtyneet1939-45/`) respectively. A previous modeling decision was that when a class containing custom properties that were not available in the CRM is needed, it will be created as a subclass of a corresponding CRM class. The full WarSampo schema is published on GitHub[10].

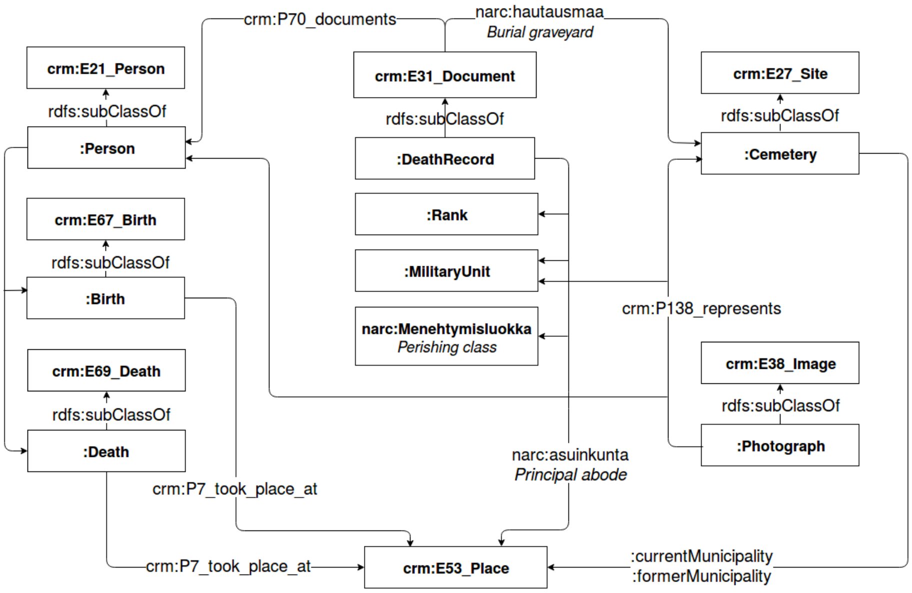


Fig. 1. The extension of WarSampo data model for cemeteries

[10] `https://github.com/SemanticComputing/Warsampo-schema`

The dataset of the Finnish war casualties consists of about 95,000 death records, and originates from the Finnish National Archives. It has served as the primary source of person instances in WarSampo. Because it was created before the WarSampo Data Service, it uses a different namespace for some properties. As can be seen from figure 1, the death records are modeled as CRM documents, which have custom properties for different pieces of information about the casualty. The death records are linked to corresponding WarSampo person instances, military units, military ranks, and wartime municipalities [10].

The death records were originally annotated with information about cemeteries, but this information was limited only to the cemetery name, and the municipality in which the cemetery is located. Because the "War Cemeteries in Finland" project and the original death record database used the same listing of war cemetery names as a starting point, it is a straightforward operation to append the new cemetery data and photographs to the existing cemetery instances by using only the name of the cemetery as a connecting link. The cemeteries provide a more stable basis for local history than the death records' temporally changing municipalities, which have already changed greatly since the war.

4 New Views to WW2 Casualties

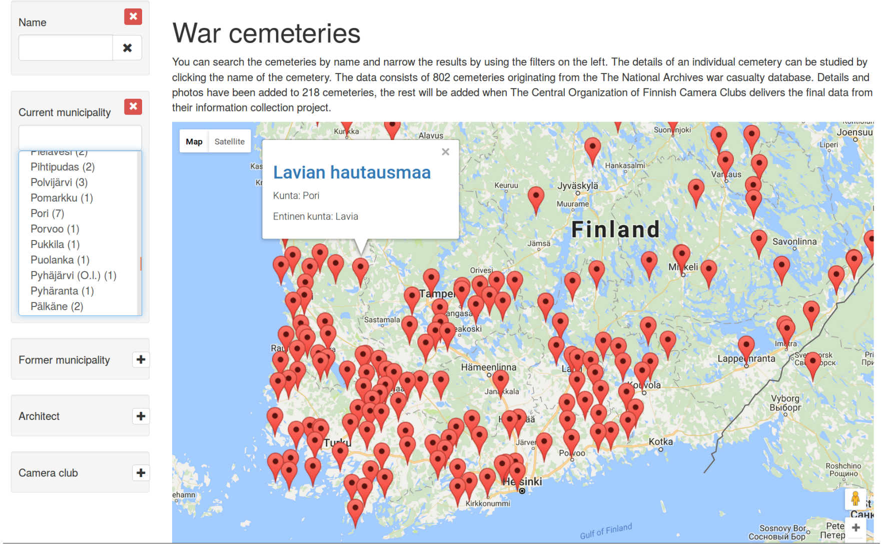
The new cemetery perspective and the new visualizations in the casualties perspective have been developed to gain new insights from the data based on the community-level aspect provided by the cemeteries. Since there is not enough data about the casualties to construct life stories of individual soldiers as biographies, the new views focus on supporting prosopographical study of the war casualties in which people are studied as groups. This approach is useful when there is not enough information about individual people to construct biographies, but the amount of individuals is large enough to study the data as groups of people using, for example, visualizations. Information visualization is a useful approach for analyzing complex spatiotemporal and multivariate data [8].

4.1 Cemetery Perspective and Information Pages

The user interface of the new Cemetery perspective[11] is presented in figure 2. The user can browse all cemeteries, or search the cemeteries by name and narrow the results by using the filters on the left. The search results can be viewed as a listing, or on a map which provides a global view of the cemeteries. When the user clicks the name of a cemetery, a cemetery information page opens.

Figure 3 illustrates the top section of a cemetery information page. The top section contains all information provided by the "War Cemeteries in Finland" project. Below that the user can browse through the photographs of the cemetery with a full screen photo gallery. As is the case with all other WarSampo information pages, the right column contains links to related resources. In this

[11] Premilinary version at `http://www.sotasampo.fi/en/cemeteries`



Name

Current municipality

Pielavesi (2)
Pihtipudas (2)
Polvijärvi (3)
Pomarkku (1)
Pori (7)
Porvoo (1)
Pukkila (1)
Puolanka (1)
Pyhäjärvi (O.l.) (1)
Pyhäranta (1)
Pälkäne (2)

Former municipality

Architect

Camera club

War cemeteries

You can search the cemeteries by name and narrow the results by using the filters on the left. The details of an individual cemetery can be studied by clicking the name of the cemetery. The data consists of 802 cemeteries originating from the The National Archives war casualty database. Details and photos have been added to 218 cemeteries, the rest will be added when The Central Organization of Finnish Camera Clubs delivers the final data from their information collection project.

Lavian hautausmaa

Kunta: Pori

Entinen kunta: Lavia

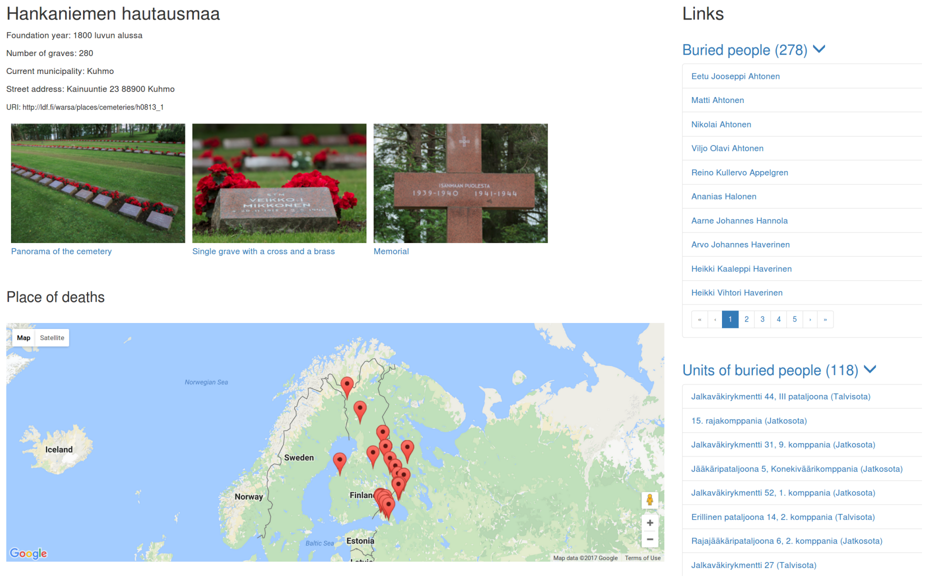
Fig. 2. Cemetery perspective

context links to buried people and their units are provided. Below the photographs there is a map visualization based on the places of death of the buried people. This example shows that the death places are concentrated on the eastern front of Finland. By clicking the map pin the user can move onwards to the person information page.

The bottom section of a cemetery information page, partly illustrated in figure 4, contains several visualizations for studying the buried people as a prosopographical group:

Distribution over units. A pie chart for showing the most common units of one's hometown soldiers.
Distribution over ranks. The most common ranks as a pie chart.
Age distribution. The ages at death as a bar chart. This visualization can also be found in the Casualties perspective with customizable options.
Distribution over causes of death. Causes of death as a pie chart.

For prosopographical research the pie chart slices and bar chart bars were made interactive, so that by clicking them the user can examine the persons that belong to the group in question and visit their information pages. On the bottom there is a link to the casualties perspective, where the user can study the buried people further as a group with a faceted search interface and newly added visualizations. A design decision originating from the needs of the user groups was to place certain visualizations to the cemetery information page, and leave the more customizable visualizations (e.g. soldier life paths) to the more general casualties perspective.



Hankaniemen hautausmaa

Foundation year: 1800 luvun alussa

Number of graves: 280

Current municipality: Kuhmo

Street address: Kainuuntie 23 88900 Kuhmo

URI: http://ldf.fi/warsa/places/cemeteries/h0813_1

Panorama of the cemetery

Single grave with a cross and a brass

Memorial

Place of deaths

Links

Buried people (278)

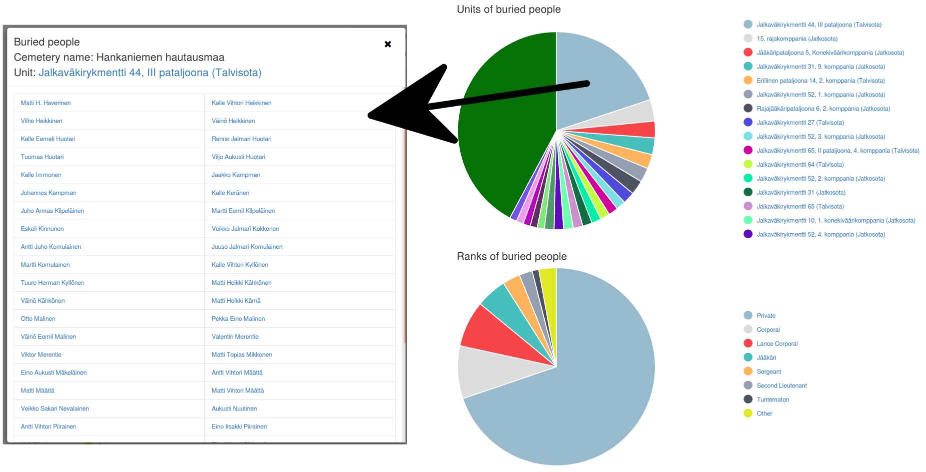
- Eetu Jooseppi Ahtonen
- Matti Ahtonen
- Nikolai Ahtonen
- Viljo Olavi Ahtonen
- Reino Kullervo Appelgren
- Ananias Halonen
- Aarne Johannes Hannola
- Arvo Johannes Haverinen
- Heikki Kaaleppi Haverinen
- Heikki Vihtori Haverinen

« ‹ 1 2 3 4 5 › »

Units of buried people (118)

- Jalkaväkirykmentti 44, III pataljoona (Talvisota)
- 15. rajakomppania (Jatkosota)
- Jalkaväkirykmentti 31, 9. komppania (Jatkosota)
- Jääkäripataljoona 5, Konekiväärikomppania (Jatkosota)
- Jalkaväkirykmentti 52, 1. komppania (Jatkosota)
- Erillinen pataljoona 14, 2. komppania (Talvisota)
- Rajajääkäripataljoona 6, 2. komppania (Jatkosota)
- Jalkaväkirykmentti 27 (Talvisota)

Fig. 3. The top section of a cemetery information page



Buried people ✖
Cemetery name: Hankaniemen hautausmaa
Unit: Jalkaväkirykmentti 44, III pataljoona (Talvisota)

Matti H. Haverinen	Kalle Vihtori Heikkinen
Vilho Heikkinen	Väinö Heikkinen
Kalle Eemeli Huotari	Renne Jalmari Huotari
Tuomas Huotari	Viljo Aukusti Huotari
Kalle Immonen	Jaakko Kampman
Johannes Kampman	Kalle Keränen
Juho Armas Kilpeläinen	Martti Eemil Kilpeläinen
Eskeli Kinnunen	Veikko Jalmari Kokkonen
Antti Juho Komulainen	Juuso Jalmari Komulainen
Martti Komulainen	Kalle Vihtori Kyllönen
Tuure Herman Kyllönen	Matti Heikki Kähkönen
Väinö Kähkönen	Matti Heikki Kärnä
Otto Malinen	Pekka Eino Malinen
Väinö Eemil Malinen	Valentin Merentie
Viktor Merentie	Matti Topias Mikkonen
Eino Aukusti Mäkeläinen	Antti Vihtori Määttä
Matti Määttä	Matti Vihtori Määttä
Veikko Sakari Nevalainen	Aukusti Nuutinen
Antti Vihtori Piirainen	Eino Iisakki Piirainen

Units of buried people
Jalkaväkirykmentti 44, III pataljoona (Talvisota)
15. rajakomppania (Jatkosota)
Jääkäripataljoona 5, Konekiväärikomppania (Jatkosota)
Jalkaväkirykmentti 31, 9. komppania (Jatkosota)
Erillinen pataljoona 14, 2. komppania (Talvisota)
Jalkaväkirykmentti 52, 1. komppania (Jatkosota)
Rajajääkäripataljoona 6, 2. komppania (Jatkosota)
Jalkaväkirykmentti 27 (Talvisota)
Jalkaväkirykmentti 52, 3. komppania (Jatkosota)
Jalkaväkirykmentti 65, II pataljoona, 4. komppania (Talvisota)
Jalkaväkirykmentti 64 (Talvisota)
Jalkaväkirykmentti 52, 2. komppania (Jatkosota)
Jalkaväkirykmentti 31 (Jatkosota)
Jalkaväkirykmentti 65 (Talvisota)
Jalkaväkirykmentti 10, 1. konekiväärikomppania (Jatkosota)
Jalkaväkirykmentti 52, 4. komppania (Jatkosota)
Ranks of buried people
Private
Corporal
Lance Corporal
Jääkäri
Sergeant
Second Lieutenant
Tuntematon
Other

Fig. 4. Visualizations for studying the buried people as a prosopographical group on a cemetery information page

4.2 Casualties Perspective Visualizations

The casualties perspective[12] of WarSampo provides a faceted search interface to the Finnish WW2 casualty data as a web page [10]. The faceted search web functionality is based on the SPARQL Faceter library [9]. The source code of the Casualties perspective in WarSampo is openly available in GitHub[13].

The faceted search results have previously been displayed in a data table, but the perspective has been extended to also support visualizing the results based on the facet selections. This provides a way to easily analyze the data by using the faceted search interface to filter the results based on what the user is interested in. For example you could select only soldiers that have been born in a certain place, or that have been buried in a certain cemetery.

After filtering the results with the facets, the user can choose from several different ways of displaying the results from the *Results Display* drop-down menu at the top of the page. The supported methods of displaying the faceted search results are:

Table. The original method of displaying the casualties as a table.

Age distribution. Age distribution of casualties as a column chart.

Soldier paths. Soldier life paths visualized as a sankey diagram. The default steps in the diagram are the municipalities of birth, residence and death, and the military cemetery. Steps can be customized by the user by selecting properties to be used for each step.

Statistics. Distribution over an arbitrary property as a bar chart. The user can select the property to use for the diagram.

Fig. 5 shows the age distribution of casualties in Hietaniemi cemetery in Helsinki, which is the military cemetery with the highest number of casualties.

Fig. 6 presents a screenshot of the soldier life paths diagram, showing the life paths of 40 soldiers. The diagram shows where the soldiers were born, where they lived, where they died, and where they are buried. The birth, residence, and death are given on a municipality level, whereas the cemetery is given as the exact cemetery. In this case the facets have been used to filter the casualties to only show persons buried in the cemetery of Inari in Ivalo. A large portion of places of death are unknown in the data, as is the case here as well. All in all 56% of the casualties in the dataset have an unknown place of death.

The faceted search makes it possible to further narrow the search down e.g. to only include casualties who served in a certain military unit. Figure 7 show the whole user interface of the Statistics visualization view. Based on the facet selections, the diagram currently shows occupations of soldiers with the military rank of private, who were married, were born in Helsinki, lived in Helsinki, and are buried in Hietaniemi war cemetery in Helsinki. The most common occupations in the group are workman, chauffeur and labourer.

[12] `http://sotasampo.fi/en/casualties`

[13] `https://github.com/SemanticComputing/WarSampo-death-records`

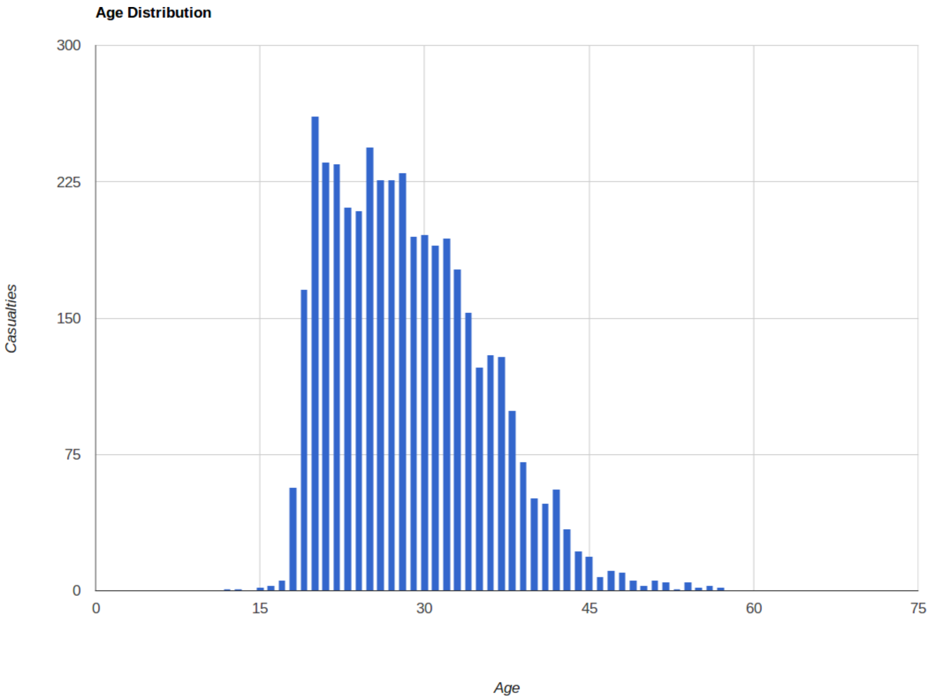


Fig. 5. Age distribution of casualties in the most ”crowded” war cemetery, Hietaniemi in Helsinki.

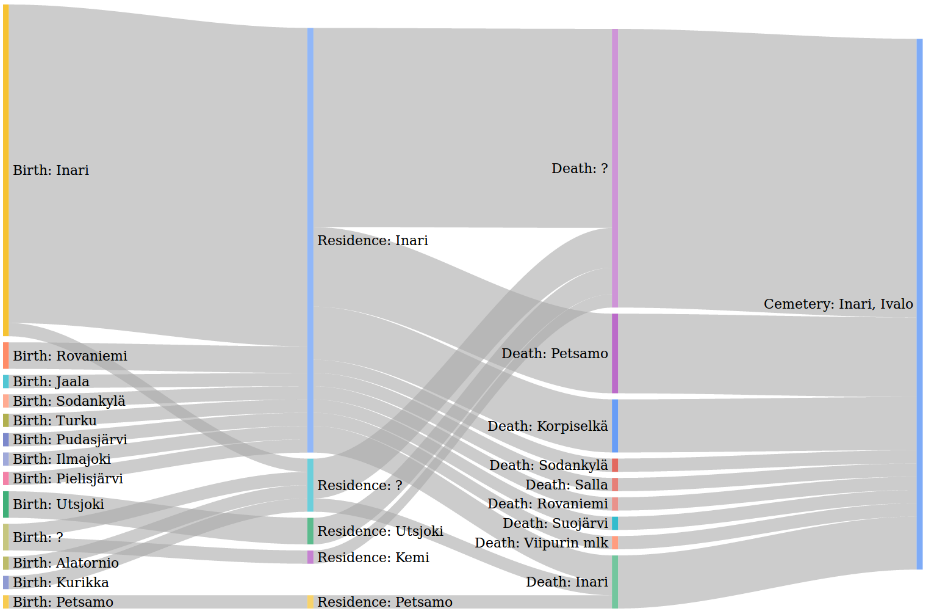
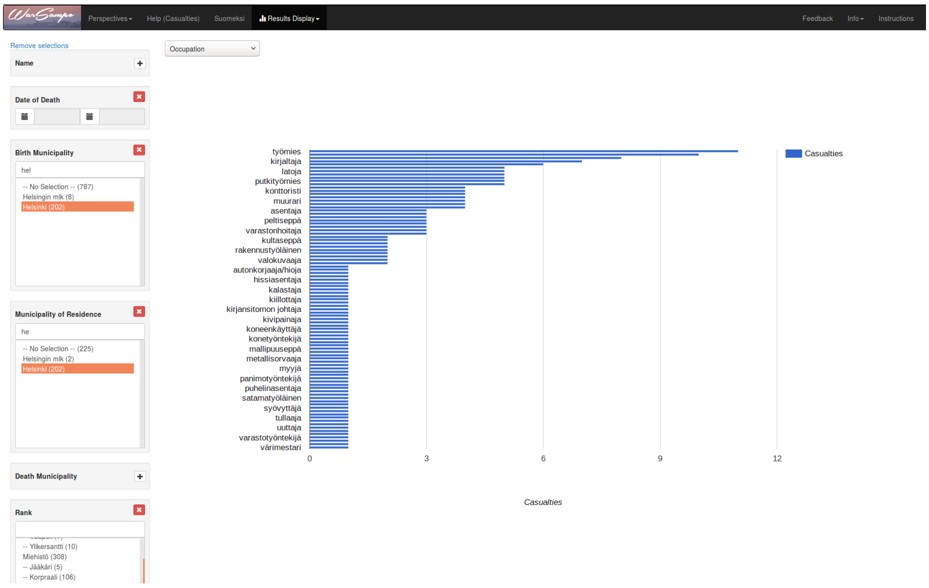


Fig. 6. Life paths of 40 soldiers buried in the cemetery of Inari in Ivalo.



Remove selections

Name +

Date of Death

Birth Municipality

hel

-- No Selection -- (787)
Helsingin mlk (8)
Helsinki (202)

Municipality of Residence

he

-- No Selection -- (225)
Helsingin mlk (2)
Helsinki (202)

Death Municipality +

Rank

-- Ylikersantti (10)
Miehistö (308)
-- Jääkäri (5)
-- Korpraali (106)

Occupation

Fig. 7. The faceted search interface and the statistics view, visualizing occupations of soldiers with private military rank, who were married, born in Helsinki, living in Helsinki, and buried in Hietaniemi war cemetery in Helsinki.

5 Related Work and Discussion

The War Graves Photographic Project[14], founded in 2008, aims to create an archive of names and photographs of all military graves and memorials from 1914 to the present day from any nationality, although the focus is on Commonwealth soldiers. Data collection is based on volunteer work, but unfortunately the data is not available through APIs or dumps, and the photographs are subject to a charge. For modeling general war related English data Historic England hosts various thesauri[15]. Also a number of previous research exists in Linked Data visualization [1,3].

This paper presented how Finnish WW2 casualty data was linked to war cemetery data that is being collected for studying the casualties from a prosopographical perspective. The WarSampo portal was extended with several new interactive views to the casualty data based on cemeteries and other community level groups. A quick comparison shows that the War Graves Photographic Project has information on 668 people buried in Hietaniemi, the biggest war cemetery in Finland, whereas WarSampo now has information on 4268 people buried there.

The Cemetery perspective and the data visualizations on cemetery information pages are useful for all user groups mentioned in section 2, but because they provide easy access and a starting point from a local point of view, private citizens might benefit most from it. These visualizations offer instant information on divisions, for example, in ranks, age groups and units. By using a local starting point for the whole casualties dataset it is also much easier to detect local irregularities in data that might be worth exploring in more detail. The new visualizations of the Casualties perspective offer more customizable views to the casualty data as a whole, which are especially useful for academic researchers.

6 Future Work

While the current tools allow for many ways to analyze and explore the data, there is still a vast amount of different kinds of visualizations that could be tried, for example many that would be based on aggregating the data in some manner. For example displaying the average amount of children in relation to some other property, like place of residence, might reveal something interesting. Also the casualty and cemetery data collected from various sources contains errors and inconsistencies. For improving the quality of the data a Linked Data tool for suggesting corrections and collecting new data from the users is in planning.

[14] `https://www.twgpp.org/`

[15] `http://thesaurus.historicengland.org.uk/`

Acknowledgements Our work is funded by the Memory Foundation for the Fallen[16], Open Science and Research Initiative[17] of the Finnish Ministry of Education and Culture, the Finnish Cultural Foundation, and the Academy of Finland.

References

1. Bikakis, N., Sellis, T.: Exploration and visualization in the web of big linked data: A survey of the state of the art. In: Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference. CEUR Workshop Proceedings, Vol-1558 (2016), http://ceur-ws.org/Vol-1558/paper28.pdf
2. 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)
3. Dadzie, A.S., Rowe, M.: Approaches to visualising Linked Data: A survey. Semantic Web 2(2), 89–124 (2011)
4. Doerr, M.: The CIDOC CRM – an ontological approach to semantic interoperability of metadata. AI Magazine 24(3), 75–92 (2003)
5. Fleming, R.: Writing biography at the edge of history. The American Historical Review 114(3), 606–614 (2009), http://dx.doi.org/10.1086/ahr.114.3.606
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., 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)
8. Kehrer, J., Hauser, H.: Visualization and visual analysis of multifaceted scientific data: A survey. IEEE transactions on visualization and computer graphics 19(3), 495–513 (2013)
9. Koho, M., Heino, E., Hyvönen, E.: SPARQL Faceter—Client-side Faceted Search Based on SPARQL. In: Troncy, R., Verborgh, R., Nixon, L., Kurz, T., Schlegel, K., Vander Sande, M. (eds.) Joint Proc. of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop. No. 1615, CEUR Workshop Proceedings (2016), http://ceur-ws.org/Vol-1615/semdevPaper5.pdf
10. 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)
11. Leskinen, J., Juutilainen, A. (eds.): Jatkosodan pikkujättiläinen. WSOY, Finland (2005)
12. Veit, R.F., Nonestied, M.: New Jersey cemeteries and tombstones: history in the landscape. Rutgers University Press (2008)
13. Verboven, K., Carlier, M., Dumolyn, J.: A short manual to the art of prosopography. In: Prosopography Approaches and Applications. A Handbook, pp. 35–70. Unit for Prosopographical Research (Linacre College) (2007), http://dx.doi.org/1854/8212

[16] http://www.muistosaatio.fi/

[17] http://openscience.fi/

Publication VI

Mikko Koho, Esko Ikkala, and Eero Hyvönen. Reassembling the Lives of Finnish Prisoners of the Second World War on the Semantic Web. Accepted for publication in *Proceedings of the Third Conference on Biographical Data in the Digital Age (BD 2019), Varna, Bulgaria*, CEUR Workshop Proceedings, 9 pages, in press, September 2019.

Reassembling the Lives of Finnish Prisoners of the Second World War on the Semantic Web

Mikko Koho[1], **Esko Ikkala**[1], **Eero Hyvönen**[1,2]

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

Abstract

This paper presents the first results of a new, ninth application perspective for the semantic portal *WarSampo – Finnish WW2 on the Semantic Web*, based on a database of ca. 4 450 Finnish prisoners of war in the Soviet Union. Our key idea is to reassemble the life of each prisoner of war by using Linked Data, based on information about the person in different data sources. Using the enriched aggregated data, a biographical global "home page" for each prisoner of war can be created, that is more complete than information in individual data sources. The application perspective is targeted to the researchers of military history, to study and analyze the data in order to form new research questions or hypotheses, as well as to public in the large looking for information, e.g., about their relatives that were captured as prisoners of war. Employing the faceted search of the application perspective, prosopographical research on subgroups of prisoners is possible.

1 Introduction

Representing biographical texts as Linked Data leads to a paradigm change in publishing biographical collections (Hyvönen et al., 2019): the lives can then not only be read as texts by humans but also be processed and analyzed by computational means (Fokkens et al., 2017; Warren et al., 2016), opening new possibilities in Digital Humanities (Gardiner and Musto, 2015) research for biography and prosopography (Verboven et al., 2007) as well as for data reuse in applications. The same idea of Linked Data can be applied also when biographical data is available in semi-structured or structured form from different data sources: the data about a person can be aggregated, harmonized, and reassembled into a global knowledge graph that gives a more complete picture of the biographee than any individual source alone. Based on the knowledge graph, a biography of the biographee can be generated or alternatively a semi-structured "home page" presenting her/his life. The latter approach was introduced in the semantic portal *WarSampo – Finnish WW2 on the Semantic Web*[1] (Hyvönen et al., 2016), a web service in use in Finland that had 230 000 users in 2018, typically looking for information about their relatives killed in action during the Second World War (WW2).

This paper presents a new, ninth application perspective *Prisoners of War* to be included in WarSampo. This perspective was created for studying individual people, documented in a new prisoners of war (POW) database, as well as groups of them for prosopographical analysis. The new data was aligned with and integrated into the WarSampo person data, which is mostly based on the Finnish WW2 casualties of war[2] database of the National Archives of Finland. The new application perspective enables studying not only individuals but also prosopographical studies of the prisoners using either the whole dataset or subsets of it based on user interest and selections in a faceted search (Tunkelang, 2009) view.

The new prisoners of war dataset was originally published as a book (Alava et al., 2003). For integrating and publishing the data as a part of WarSampo, it has been further extended, cleaned, and validated by domain experts using, e.g., information from many war-time archives in Finland and Russia. This paper builds on previous work on WarSampo, which has discussed the Linked Data publication and data model (Koho et al., 2018a), and the data integration challenges (Koho et al., 2018b). Reconstructing the biographies of the casualties of war in WarSampo has been previously presented in (Koho et al., 2017). In contrast to the casualties of war dataset, the POW register can have multiple values for a single property, and contains sources of information for individual data values, creating a need for handling conflicting information about a person.

In the following, the underlying data model and data production process is first explained. After this, the main functionalities of the application from an end user perspective are explained, as well as the technical implementation. In conclusion, the contributions of the work are summarized and contrasted with related work.

2 Data Model and Data

The prisoners dataset consists mainly of a register of the Finnish prisoners of war in WW2, containing a spreadsheet of about 4 450 soldiers, auxiliary forces, and civilians captured by the army of the Soviet Union. Additional spreadsheets contain information about POW camps and hospitals, as well as the primary data sources. The data includes

[1] This semantic portal was released in 2015 and is in use at https://sotasampo.fi/en/. More information about the project is available at home page https://seco.cs.aalto.fi/projects/sotasampo/en/.

[2] http://kronos.narc.fi/menehtyneet/

also separate documents about the prisoners of war to provide additional information, such as video interviews, images and archived documents.

The original information sources are mostly various registers in Finnish and Russian archives (Alava et al., 2003). Information in different sources can be contradictory, hence it is important to preserve the data source for each individual piece of information. A formatting was agreed upon to allow multiple values with source information already in the original spreadsheet that the domain experts worked on. The data formatting evolved as a collaboration between the domain experts maintaining the original dataset, and the WarSampo team of Linked Data experts. Also other agreements on the spreadsheet structure were needed: 1) separation and cleaning of values that will be linked to the WarSampo domain ontologies, 2) local identifiers for entities that are used in multiple spreadsheets, and 3) how to express partially or completely missing information.

The WarSampo infrastructure, data service, and semantic portal was chosen as the primary data publication platform by the stakeholders, which include the National Archives of Finland, and the Association for Cherishing the Memory of the Dead of the War.

21 Prisoners of War as Linked Data

The WarSampo *Linked Open Data infrastructure* is built to support integrating new datasets into WarSampo, by extending both the data model and the data content. The data is published openly online for everyone to use. The WarSampo web portal then provides different perspectives to the interlinked datasets, as customized web applications. New perspectives can be added to provide views to new datasets, or to show new features of the existing data.

In *Linked Data* (Heath and Bizer, 2011), information is presented as *RDF* graphs and all resources in the data have unique identifiers. This enables identifying and sharing common resources, e.g. people, places, and military ranks between the datasets, thus creating an interlinked *knowledge graph*.

A simple primary data model is used for the prisoner records, in which one prisoner record corresponds to one row in the source spreadsheet, with each column mapped to a distinct property. So all of the personal information about each captured individual is contained in the prisoner record, resembling the data model of the WarSampo death records (Koho et al., 2017). The properties and classes of prisoner records and death records have been harmonized using the dumb-down principle of Dublin Core[3], i.e., by using shared super-properties and super-classes where applicable. By mapping columns directly to properties, the data can be shown to the end user in an intuitive way, resembling the original spreadsheet.

WarSampo uses the *CIDOC Conceptual Reference Model (CRM)*[4] as the harmonizing data model. Prisoner records are modeled as instances of the CRM document class `E31_Document`. In addition, this data is then used to create CIDOC CRM descriptions of the actual people and events, when appropriate. WarSampo person instances (Leskinen et al., 2017) in the actor ontology are enriched using the prisoner records. New person instances are created for people that do not already exist in the ontology, which is the case for most of the war prisoners. The prisoner records then document the person instance through the CRM property `P70_documents`. The full WarSampo data model is published on GitHub[5].

22 Data Conversion

It has been understood from our previous work, that the data transformations need to be repeatable, automated processes (Koho et al., 2018b), in the dynamic infrastructure where there is frequently a need to adapt to changes. An automatic data processing pipeline[6] was developed to integrate the POW data into WarSampo linked data infrastructure. The pipeline handles data transformation, validation, linking, and harmonization.

The pipeline transforms the spreadsheets into RDF, mapping the spreadsheet columns to RDF properties, with possibly multiple values per property, and containing annotations for primary information sources. Automatic probabilistic entity linking processes then link the records to the WarSampo domain ontologies of military ranks, units, occupations, people, and places. Original literal values are also retained as separate properties.

The original POW register is maintained in spreadsheet format, which can be easily integrated into WarSampo with our automated transformation process when the spreadsheet is updated, provided that the structure stays the same. Also if the linked domain ontologies are updated, the whole integration process can be redone to account for the changes in the probabilistic entity linking.

The cell formatting is validated during the data transformation process. Also other simple data validation rules are applied to find anomalies during data conversions. The validation reports help the domain experts to improve the quality of the source data.

Some parts of the data had to be left out of the online data publication due to privacy issues. This is done automatically based on the date when a person has died. If there is no information about an individual's date of death, it is assumed that they may still be alive, and their personal information, including given names, is removed, effectively pseudonymizing them. For prisoners who are known to have died less than 50 years ago, health related information is removed, based on the columns of the original spreadsheet that might contain health related information.

23 Interlinking within WarSampo

Matching the people in the prisoner records to the ca. 100 000 people already existing in the WarSampo actor ontology is one of the most challenging aspects of the data transformation pipeline. The data model and contents are

[3] `http://dublincore.org/usage/documents/principles/`

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

[5] `https://github.com/SemanticComputing/Warsampo-schema`

[6] Source codes for data conversion and linking are available online: `https://github.com/SemanticComputing/WarPrisoners`.

different, and many pieces of personal information can be missing on both sides. In the first results of the person linking, we were able to link 1431 prisoner records to existing WarSampo person instances, corresponding to 32% of all prisoner records (Koho et al., 2018a). The person linking uses probabilistic *record linkage* (Gu et al., 2003; Gregg and Eder, 2019) (aka. *deduplication*) with a machine learning approach, in which each POW's information is compared with the information in the WarSampo person instances to find matches that have high enough similarity. Initially the record linkage value comparisons were weighted based on domain knowledge, which was then iterated for better accuracy, and finally a manually curated list of matches was taken to serve as training data for the machine learning approach. The machine learning approach can adapt to data changes on both sides in the record linkage, without having to manually inspect the linking results and adjust the weights.

New person instances are created from the unlinked prisoner records and added into the actor ontology. With the probabilistic record linkage, it is possible that a record is not mapped simply because there is not enough information about either the POW record, or the person instance, to create a mapping between them. Modifying the information in either the POW data or in the actor ontology means that the whole record linkage process should be redone.

Other information is also linked to WarSampo domain ontologies. Of military ranks, 99% were linked to the WarSampo military ranks domain ontology. Of military units, 91% were linked to pre-existing military units in the actor ontology.

Domain ontologies differ from each other by nature. For example, covering and disambiguating all military ranks is clearly a simpler task than performing the same task with all wartime places. In general, it is not realistic to assume that the domain ontologies completely cover their domain.

Other information still to be linked to WarSampo domain ontologies are war-time municipalities. More accurate place information could also be linked, but due to the ambiguous nature of the names, this would lead to a high level of error, based on initial experiments.

The created Linked Data stores source information when present in the original data. There are many ways of presenting this kind of provenance information in RDF (Hartig, 2009; Zhao et al., 2010). The approach used with the prisoners of war dataset is storing source information using RDF reification with the DCMI Metadata Terms[7] property *source*.

24 Biographical Data

Each person's basic personal information in the dataset contains columns like first and last names, dates of birth, return from captivity, and death, municipality of birth, domicile and death, and occupation, marital status, and number of children. These enable building some understanding about the life of the person before the war, and in case of survivors, also after the war.

Structured information is also gathered of the events of going missing and being captured, like the place and time. Biographically interesting information is also given as prose about being captured, the cause of death and burial place, and other information. These all are structured to contain the information source, and can often contain different pieces of information from different sources. Information on confiscated possessions and their estimated value sheds light to what kind of valuable personal possessions a person had. Information is also given about the occurrence of a person in Soviet war propaganda magazines or fliers, either in pictures or text.

3 Prisoners of War in the WarSampo Portal

A new application perspective was created into the WarSampo portal for studying, exploring and analyzing the prisoners of war dataset as a whole. Also the existing Warsampo Persons perspective, which generates a “home page” for each person in the WarSampo knowledge graph, was extended to show possibly contradictory data originating from multiple sources (e.g. death records, prisoner records, Wikipedia). The Prisoner perspective application is open-source, and available online[8].

31 Biographical View in the Persons perspective

The WarSampo Persons perspective offers a general search of people in the WarSampo knowledge graph. Each person is provided with a biographical view, a home page, that reassembles the biographical knowledge of the person from the WarSampo datasets, into a structured format.

Figure 1 shows an example of a soldier's home page, where the information is combined from a prisoner record and a death record. The left side of the page contains a person selector and a text box for filtering the people by name. The details of a selected person are displayed on the right. Information usually exists from birth to death, with a clear and understandable focus on the war-time events. A property (e.g. occupation) may contain multiple values. In order to make the biographical view as transparent as possible, all values have been supplemented with a reference to the information source. In the figure, source number 2 refers to the POW register. There is a total of 12 sources of information for the particular person, which includes also a death record, and 10 different sources from the POW register.

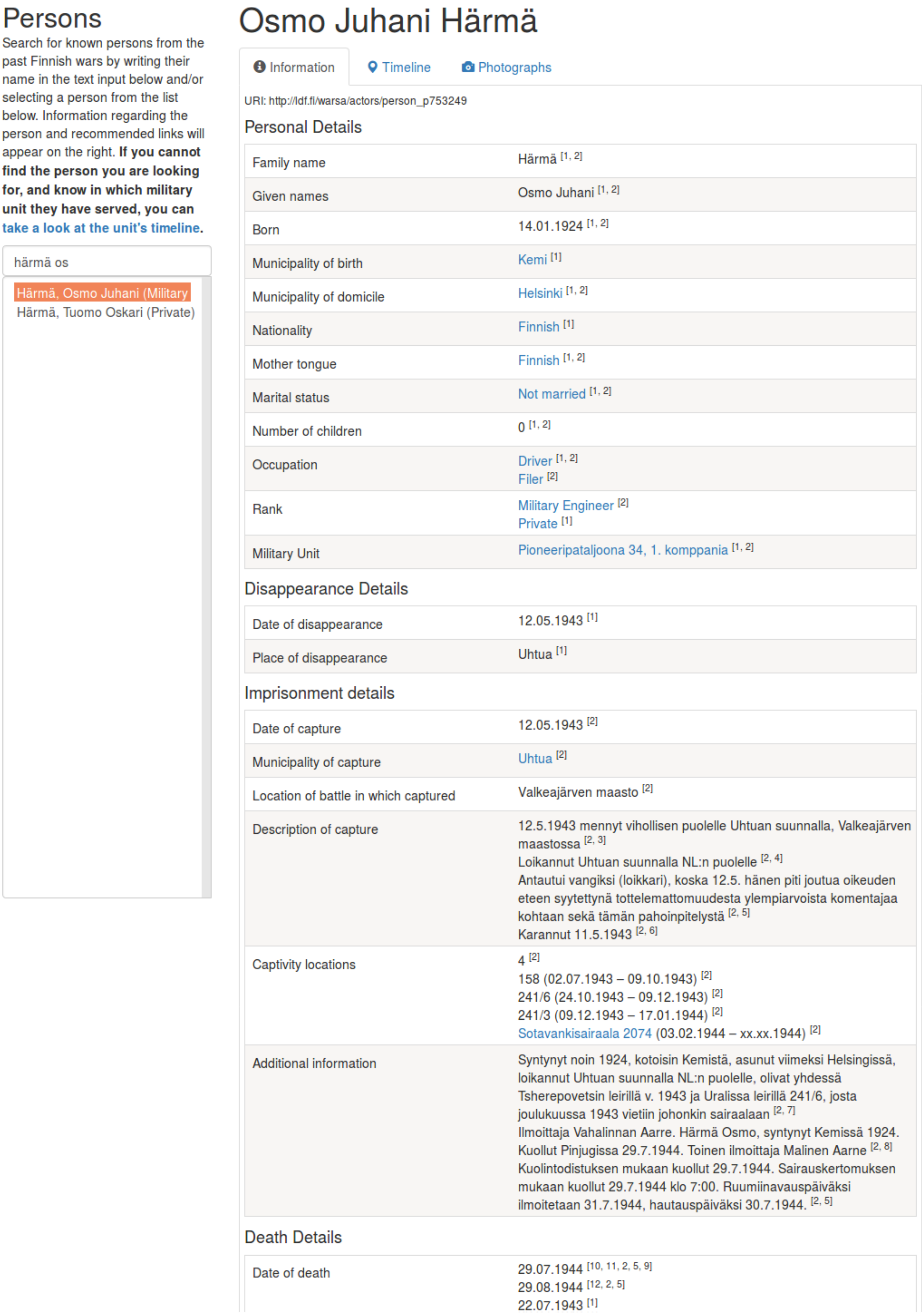
The values that have been linked to WarSampo domain ontologies are shown as links to corresponding home pages. The idea here is that the WarSampo semantic portal acts as a customized graphical RDF browser, which makes it possible for the user to find surprising connections between the individual resources of the WarSampo knowledge graph.

32 Prosopographical Prisoners Perspective

The Prisoners perspective is based on the previously released Casualties perspective (Koho et al., 2017). The main design principle of these perspectives is to target one core class of WarSampo knowledge graph (e.g., prisoner record) and provide the user with a faceted search (Tunkelang,

[7] http://dublincore.org/documents/dcmi-terms/

[8] https://github.com/SemanticComputing/prisoners-demo



Persons

Search for known persons from the past Finnish wars by writing their name in the text input below and/or selecting a person from the list below. Information regarding the person and recommended links will appear on the right. **If you cannot find the person you are looking for, and know in which military unit they have served, you can take a look at the unit's timeline.**

härmä os

Härmä, Osmo Juhani (Military
Härmä, Tuomo Oskari (Private)

Osmo Juhani Härmä

URI: http://ldf.fi/warsa/actors/person_p753249

Personal Details

Family name	Härmä [1, 2]
Given names	Osmo Juhani [1, 2]
Born	14.01.1924 [1, 2]
Municipality of birth	Kemi [1]
Municipality of domicile	Helsinki [1, 2]
Nationality	Finnish [1]
Mother tongue	Finnish [1, 2]
Marital status	Not married [1, 2]
Number of children	0 [1, 2]
Occupation	Driver [1, 2] Filer [2]
Rank	Military Engineer [2] Private [1]
Military Unit	Pioneeripataljoona 34, 1. komppania [1, 2]

Disappearance Details

Date of disappearance	12.05.1943 [1]
Place of disappearance	Uhtua [1]

Imprisonment details

Date of capture	12.05.1943 [2]
Municipality of capture	Uhtua [2]
Location of battle in which captured	Valkeajärven maasto [2]
Description of capture	12.5.1943 mennyt vihollisen puolelle Uhtuan suunnalla, Valkeajärven maastossa [2, 3] Loikannut Uhtuan suunnalla NL:n puolelle [2, 4] Antautui vangiksi (loikkari), koska 12.5. hänen piti joutua oikeuden eteen syytettynä tottelemattomuudesta ylempiarvoista komentajaa kohtaan sekä tämän pahoinpitelystä [2, 5] Karannut 11.5.1943 [2, 6]
Captivity locations	4 [2] 158 (02.07.1943 – 09.10.1943) [2] 241/6 (24.10.1943 – 09.12.1943) [2] 241/3 (09.12.1943 – 17.01.1944) [2] Sotavankisairaala 2074 (03.02.1944 – xx.xx.1944) [2]
Additional information	Syntynyt noin 1924, kotoisin Kemistä, asunut viimeksi Helsingissä, loikannut Uhtuan suunnalla NL:n puolelle, olivat yhdessä Tsherepovetsin leirillä v. 1943 ja Uralissa leirillä 241/6, josta joulukuussa 1943 vietiin johonkin sairaalaan [2, 7] Ilmoittaja Vahalinnan Aarre. Härmä Osmo, syntynyt Kemissä 1924. Kuollut Pinjugissa 29.7.1944. Toinen ilmoittaja Malinen Aarne [2, 8] Kuolintodistuksen mukaan kuollut 29.7.1944. Sairauskertomuksen mukaan kuollut 29.7.1944 klo 7:00. Ruumiinavauspäiväksi ilmoitetaan 31.7.1944, hautauspäiväksi 30.7.1944. [2, 5]

Death Details

Date of death	29.07.1944 [10, 11, 2, 5, 9] 29.08.1944 [12, 2, 5] 22.07.1943 [1]

Figure 1: The Persons perspective showing part of a person's home page.

2009; Oren et al., 2006) interface, which initially renders a result set that contains all instances of the target class as a paginated table. This way we ease off the "blank search field problem", where a new user does not know what kind of query terms should be used for meaningful results. The initial result set can be narrowed down by using various facets (e.g., military unit or prison camp).

Figure 2 shows a part of the Prisoners perspective user interface. Facets are presented on the left of the user interface. The number of hits (instances of the target class) produced by each facet value is calculated dynamically and is shown in parenthesis. Facet values leading to an empty result set are hidden. To reduce unnecessary data fetching, most of the facets are disabled by default. They can be activated by clicking the plus sign on the facet header. The facets are *name, date of being captured as a POW, date of death, military unit, military rank, POW camps where the person has been, occupation, marital status, number of children, birth municipality, place of being captured, and place of death.*

The results are displayed on the right side of the user interface. The result set, based on the facet selections, can be shown as a table, or shown with three different visualizations:

1. a distribution chart over a selected property, with property choices: military rank, military unit, occupation, number of children, birth municipality, municipality of residence, place of being captured, and place of death,
2. an age distribution chart at the time of capturing,
3. a sankey diagram of soldier life paths based on known geographical locations at different times, starting from the municipality of birth, and ending to the municipality of death.

The results display mode can be selected using the button in the top bar. In Figure 2, the results are displayed as a table, with each row corresponding to a single prisoner record, with several key properties mapped to separate columns.

Figure 3 shows the age distribution of all soldiers whose rank is private at the time when they have been captured as a prisoner of war. Figure 4 shows the military rank distribution of the soldiers that were born in Helsinki.

The common usage scenario of the average user is to search for information about their relatives who have participated in the war. This can be achieved most easily with the table view of results and using the different facets, and mostly the name facet, where a person can search with just a part of the name to get all the results containing that. Another way to find relatives, who historically are often situated in the same region, is to filter the results with the birth municipality facet.

Another usage scenario is studying and analyzing the data by a historian or an interested citizen. The facets already provide distributions of the facet values, with the number of hits after each value. When a selection is made in one of the facets, all of the facets are updated to show the distribution of values with that selection. Further analysis can be done with the various visualizations of the facet results. New visualizations, e.g. locations of the POW camps on a map, can be added rather easily to the application, and the existing ones extended as needed.

4 Implementation

The Prisoners perspective is an AngularJS[9] web application, which consists of several modules. The facet functionality is implemented using SPARQL Faceter[10] (Koho et al., 2016), a module that provides

- a set of directives that work as configurable facets,
- a service that synchronizes the facet selections,
- a service for updating the URL parameters based on facet selections, and retrieving the facet values from URL parameters,
- a service for retrieving SPARQL results based on the facet selections, using a configurable query template.

For querying the SPARQL endpoint, mapping the SPARQL results into JavaScript objects and paging the results, we have developed another general module[11] that is being used across the WarSampo semantic portal.

In addition to the default paginated table result view, powered by the ngTable[12] directive, we have implemented several reusable visualization directives for displaying the results on modern or historical maps or as statistical distributions. For the Prisoners perspective, a new sankey visualization directive was built using Google Charts.[13]

The Persons perspective is part of the WarSampo portal AngularJS core infrastructure [14]. It was extended to fetch data to the person's homepage from the prisoner records, along with the source reifications. The page was redesigned and restructured to be able to integrate the data from the prisoner records, and to show the prisoner record data along with the information from a person instance and a death record, of which the latter may or may not be present. Showing and numbering the information sources was also a new addition.

5 Discussion

This paper presented first results of publishing the prisoners of war dataset as part of WarSampo. The POW data contains sensitive information about the individual citizens, some of whom are still alive. The publication of the data has been delayed due to the evaluation as to what information can be legally published about the individuals, and what needs to be hidden. The dataset and new portal is expected to be finally published in November 2019.

The combination of faceted search and various result visualization components forms the base of the user interface

[9] `https://angularjs.org/`

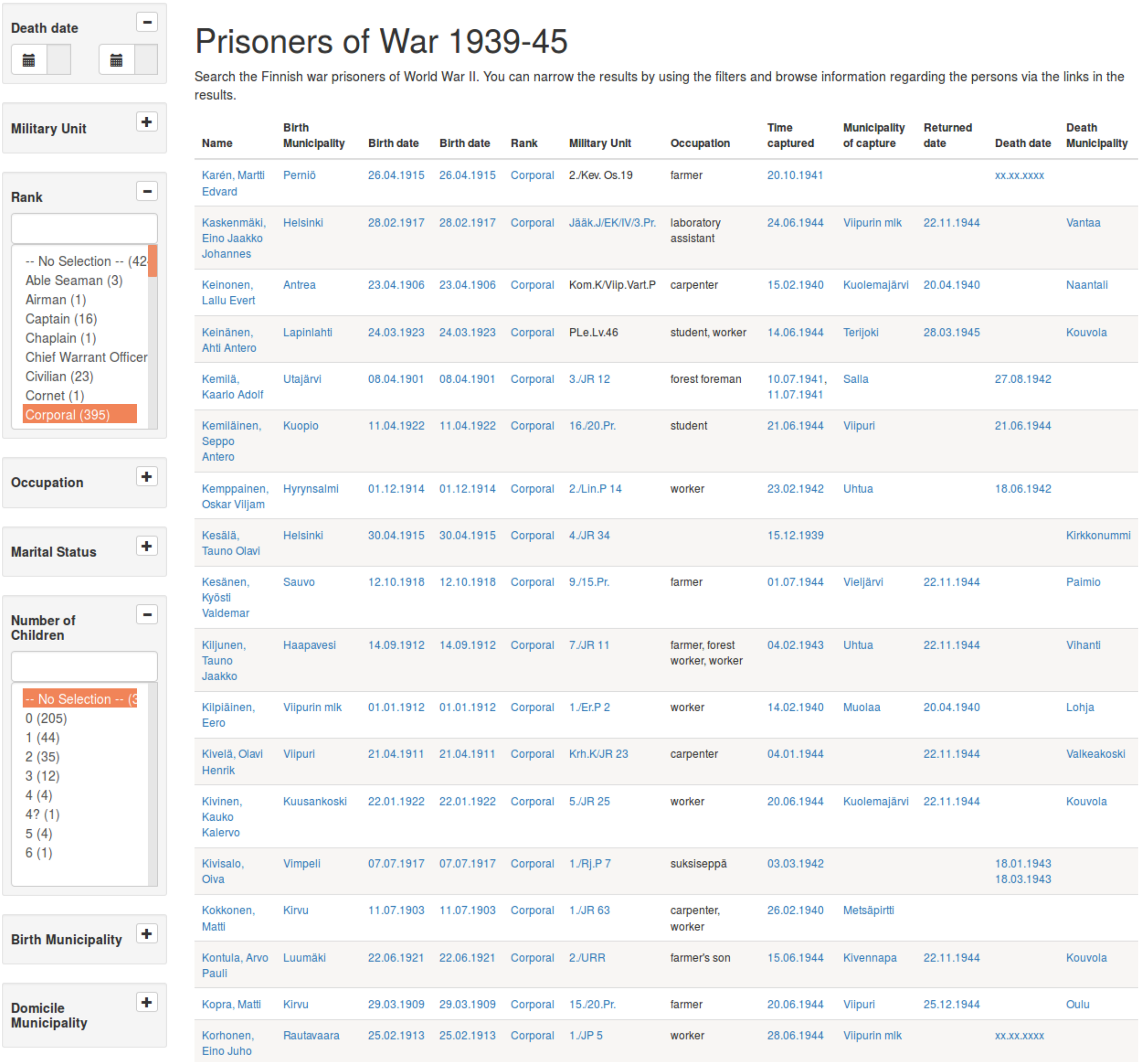
[10] `https://github.com/SemanticComputing/angular-semantic-faceted-search`

[11] `https://github.com/SemanticComputing/angular-paging-sparql-service`

[12] `https://github.com/esvit/ng-table`

[13] `https://github.com/angular-google-chart/angular-google-chart`

[14] `https://github.com/SemanticComputing/warsampo-angular-app`



Death date

Military Unit

Rank

- -- No Selection -- (42
- Able Seaman (3)
- Airman (1)
- Captain (16)
- Chaplain (1)
- Chief Warrant Officer
- Civilian (23)
- Cornet (1)
- Corporal (395)

Occupation

Marital Status

Number of Children

- -- No Selection -- (3
- 0 (205)
- 1 (44)
- 2 (35)
- 3 (12)
- 4 (4)
- 4? (1)
- 5 (4)
- 6 (1)

Birth Municipality

Domicile Municipality

Prisoners of War 1939-45

Search the Finnish war prisoners of World War II. You can narrow the results by using the filters and browse information regarding the persons via the links in the results.

Name	Birth Municipality	Birth date	Birth date	Rank	Military Unit	Occupation	Time captured	Municipality of capture	Returned date	Death date	Death Municipality
Karén, Martti Edvard	Perniö	26.04.1915	26.04.1915	Corporal	2./Kev. Os.19	farmer	20.10.1941			xx.xx.xxxx	
Kaskenmäki, Eino Jaakko Johannes	Helsinki	28.02.1917	28.02.1917	Corporal	Jääk.J/EK/IV/3.Pr.	laboratory assistant	24.06.1944	Viipurin mlk	22.11.1944		Vantaa
Keinonen, Lallu Evert	Antrea	23.04.1906	23.04.1906	Corporal	Kom.K/Viip.Vart.P	carpenter	15.02.1940	Kuolemajärvi	20.04.1940		Naantali
Keinänen, Ahti Antero	Lapinlahti	24.03.1923	24.03.1923	Corporal	PLe.Lv.46	student, worker	14.06.1944	Terijoki	28.03.1945		Kouvola
Kemilä, Kaarlo Adolf	Utajärvi	08.04.1901	08.04.1901	Corporal	3./JR 12	forest foreman	10.07.1941, 11.07.1941	Salla		27.08.1942	
Kemiläinen, Seppo Antero	Kuopio	11.04.1922	11.04.1922	Corporal	16./20.Pr.	student	21.06.1944	Viipuri		21.06.1944	
Kemppainen, Oskar Viljam	Hyrynsalmi	01.12.1914	01.12.1914	Corporal	2./Lin.P 14	worker	23.02.1942	Uhtua		18.06.1942	
Kesälä, Tauno Olavi	Helsinki	30.04.1915	30.04.1915	Corporal	4./JR 34		15.12.1939				Kirkkonummi
Kesänen, Kyösti Valdemar	Sauvo	12.10.1918	12.10.1918	Corporal	9./15.Pr.	farmer	01.07.1944	Vieljärvi	22.11.1944		Paimio
Kiljunen, Tauno Jaakko	Haapavesi	14.09.1912	14.09.1912	Corporal	7./JR 11	farmer, forest worker, worker	04.02.1943	Uhtua	22.11.1944		Vihanti
Kilpiäinen, Eero	Viipurin mlk	01.01.1912	01.01.1912	Corporal	1./Er.P 2	worker	14.02.1940	Muolaa	20.04.1940		Lohja
Kivelä, Olavi Henrik	Viipuri	21.04.1911	21.04.1911	Corporal	Krh.K/JR 23	carpenter	04.01.1944		22.11.1944		Valkeakoski
Kivinen, Kauko Kalervo	Kuusankoski	22.01.1922	22.01.1922	Corporal	5./JR 25	worker	20.06.1944	Kuolemajärvi	22.11.1944		Kouvola
Kivisalo, Oiva	Vimpeli	07.07.1917	07.07.1917	Corporal	1./Rj.P 7	suksiseppä	03.03.1942			18.01.1943 18.03.1943	
Kokkonen, Matti	Kirvu	11.07.1903	11.07.1903	Corporal	1./JR 63	carpenter, worker	26.02.1940	Metsäpirtti			
Kontula, Arvo Pauli	Luumäki	22.06.1921	22.06.1921	Corporal	2./URR	farmer's son	15.06.1944	Kivennapa	22.11.1944		Kouvola
Kopra, Matti	Kirvu	29.03.1909	29.03.1909	Corporal	15./20.Pr.	farmer	20.06.1944	Viipuri	25.12.1944		Oulu
Korhonen, Eino Juho	Rautavaara	25.02.1913	25.02.1913	Corporal	1./JP 5	worker	28.06.1944	Viipurin mlk		xx.xx.xxxx	

Figure 2: Prisoners perspective: facet selection results shown as a table view.

of the Prisoners perspective. This design has proved to be broadly applicable to many kinds of datasets. By browsing through the facets, the user can quickly see what kind of values have been used for different properties. This often reveals inconsistencies and spelling errors, if the property values have not been systemically entered or harmonized, or they are completely missing for a large number of resources. For estimating the completeness and the reliability of the dataset, looking at the actual property values is often more important than focusing on data modeling details.

Maintaining interlinked datasets and domain ontologies present new challenges (Auer et al., 2012; Maedche et al., 2003), as changes is one part need to be accounted for in other interlinked parts. The Linked Data environment is not yet mature enough to have easy-to-use tools for non-technical people to use for editing and maintaining interlinked data. Hence, the POW data is still maintained using the spreadsheet with agreed upon formatting and structuring, which can then be re-integrated easily into WarSampo. The Linked Data approach requires tighter co-operation with the domain experts and data publishers, especially in the creation phase of historical information (Boonstra et al., 2004), than more traditional data publishing ways. However, it is possible using Linked Data to create an understanding about the whole of the war, by combining information from several datasets together, which would not be easy by studying the individual datasets directly.
The historical occupations in the WarSampo datasets have recently been harmonized into a manually curated SKOS-based [15] ontology *AMMO* (Koho et al., 2019), to which the prisoner records are linked. The ontology combines synonymous occupational labels into harmonized occupation resources, and provides structures of social stratification and occupational groups. It will enable studying the prisoner records using new facets in the future, such as social

[15] https://www.w3.org/TR/skos-primer/

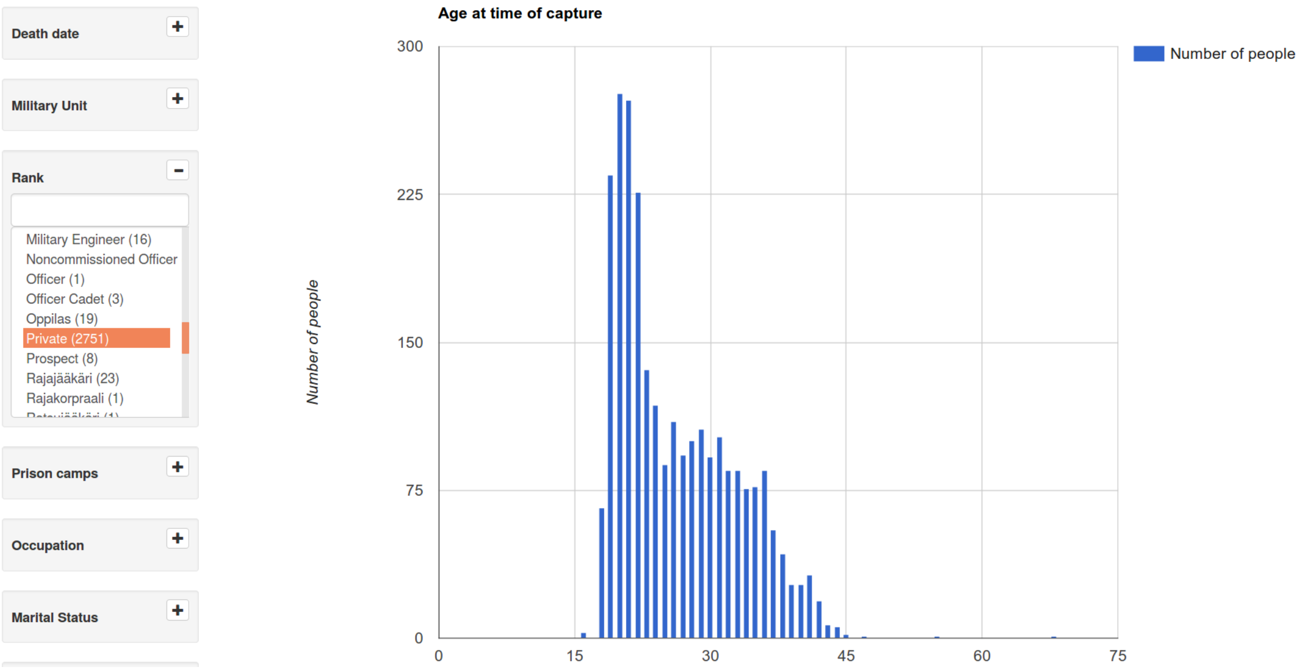


Figure 3: Prisoners perspective: age distribution of the soldiers with the military rank private.

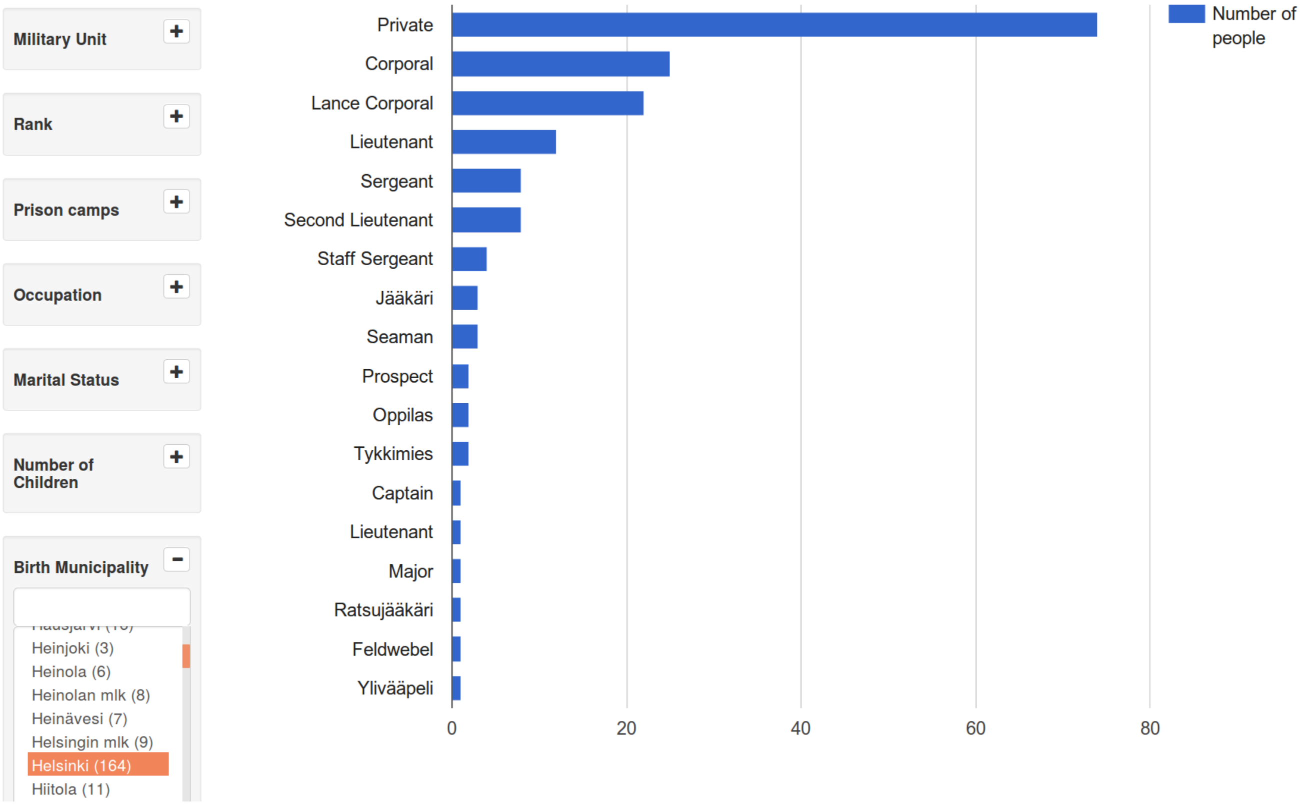


Figure 4: Prisoners perspective: statistics view

class and field of work, and facilitate the use of the dataset to answer new kinds of research questions of collaborating historians.

Integration of videos and other documents relating to the prisoners of war, will be implemented later, and will consist of expressing the document metadata in terms of CIDOC CRM, and linking the prisoners to the related document resources, which in turn contain URL links to the document files.

Integrating data into a Linked Data infrastructure is more laborious than simpler ways of publishing the data as an independent data object, which does not communicate with other datasets. However, the result of the integration is an interlinked knowledge base, where the interlinked graphs enrich each other, creating a whole that is greater than the sum of its parts (Hyvönen, 2012).

Acknowledgements

Reijo Nikkilä, Tiia Moilanen, and Pertti Suominen of The National Prisoners of War Project worked on the data as domain experts. Katri Miettinen indexed related documents for linking with persons.

Our work was funded by the Association for Cherishing the Memory of the Dead of the War[16], Teri-Säätiö, Open Science and Research Initiative[17] of the Finnish Ministry of Education and Culture, the Finnish Cultural Foundation, and the Academy of Finland.

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

6 References

Teuvo Alava, Dmitri Frolov, and Reijo Nikkilä. 2003. Rukiver. *Suomalaiset sotavangit Neuvostoliitossa. Helsinki: Edita.*

Sören Auer, Theodore Dalamagas, Helen Parkinson, and Bancilhon and. 2012. Diachronic linked data: towards long-term preservation of structured interrelated information. In *Proceedings of the First International Workshop on Open Data*, pages 31–39. ACM.

Onno Boonstra, Leen Breure, and Peter Doorn. 2004. Past, present and future of historical information science. *Historical Social Research*, 29(2):4–132.

Antske Fokkens, Serge ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, Guus Schreiber, and Victor de Boer. 2017. Biographynet: Extracting relations between people and events. In *Europa baut auf Biographien*, pages 193–224. New Academic Press, Wien.

Eileen Gardiner and Ronald G. Musto. 2015. *The Digital Humanities: A Primer for Students and Scholars*. Cambridge University Press, New York, NY, USA.

Forest Gregg and Derek Eder. 2019. Dedupe. https://github.com/dedupeio/dedupe.

Lifang Gu, Rohan Baxter, Deanne Vickers, and Chris Rainsford. 2003. Record linkage: Current practice and future directions. *CSIRO Mathematical and Information Sciences Technical Report*, 3:83.

Olaf Hartig. 2009. Provenance information in the web of data. In *Proceedings of the WWW2009 Workshop on Linked Data on the Web*, volume 538 of *CEUR Workshop Proceedings*.

Tom Heath and Christian Bizer. 2011. *Linked Data: Evolving the web into a global data space*. Synthesis Lectures on The Semantic Web: Theory and Technology. Morgan & Claypool Publishers, Palo Alto, USA.

Eero Hyvönen, Erkki Heino, Petri Leskinen, Esko Ikkala, Mikko Koho, Minna Tamper, Jouni Tuominen, and Eetu Mäkelä. 2016. 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)*, pages 758–773. Springer-Verlag.

Eero Hyvönen, Petri Leskinen, Minna Tamper, Heikki Rantala, Esko Ikkala, Jouni Tuominen, and Kirsi Keravuori. 2019. Biographysampo – publishing and enriching biographies on the semantic web for digital humanities research. In *Proceedings of the 16th Extendwed Semantic Web Conference (ESWC 2019)*, pages 574–589. Springer-Verlag.

Eero Hyvönen. 2012. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool Publishers, Palo Alto, USA.

Mikko Koho, Erkki Heino, and Eero Hyvönen. 2016. SPARQL Faceter—Client-side Faceted Search Based on SPARQL. In *Joint Proc. of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop*, number 1615. CEUR Workshop Proceedings.

Mikko Koho, Eero Hyvönen, Erkki Heino, Jouni Tuominen, Petri Leskinen, and Eetu Mäkelä. 2017. Linked death—representing, publishing, and using Second World War death records as linked open data. In Eva Blomqvist, Katja Hose, Heiko Paulheim, Agnieszka Ławrynowicz, Fabio Ciravegna, and Olaf Hartig, editors, *The Semantic Web: ESWC 2017 Satellite Events*, pages 369–383. Springer-Verlag.

Mikko Koho, Erkki Heino, Esko Ikkala, Eero Hyvönen, Reijo Nikkilä, Tiia Moilanen, Katri Miettinen, and Pertti Suominen. 2018a. Integrating prisoners of war dataset into the WarSampo linked data infrastructure. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018)*. CEUR Workshop Proceedings, March. Vol 2084.

Mikko Koho, Esko Ikkala, Erkki Heino, and Eero Hyvönen. 2018b. 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*, volume 11196. Springer-Verlag, October-November.

Mikko Koho, Lia Gasbarra, Jouni Tuominen, Heikki Rantala, Ilkka Jokipii, and Eero Hyvönen. 2019. AMMO Ontology of Finnish Historical Occupations. In *Proceedings of the The First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH'19)*, volume 2375. CEUR Workshop Proceedings, June.

Petri Leskinen, Mikko Koho, Erkki Heino, Minna Tamper, Esko Ikkala, Jouni Tuominen, Eetu Mäkelä, and Eero Hyvönen. 2017. 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, October.

Alexander Maedche, Boris Motik, Ljiljana Stojanovic, Rudi Studer, and Raphael Volz. 2003. An infrastructure for searching, reusing and evolving distributed ontologies. In *Proc. of the twelfth international conference on World Wide Web*, pages 439–448. ACM Press.

Eyal Oren, Renaud Delbru, and Stefan Decker. 2006. Extending faceted navigation for RDF data. In *International semantic web conference*, pages 559–572. Springer–Verlag.

[16] http://www.sotavainajat.net/in_english

[17] http://openscience.fi/

Daniel Tunkelang. 2009. *Faceted search*. Synthesis lectures on information concepts, retrieval, and services. Morgan & Claypool Publishers.

Koenraad Verboven, Myriam Carlier, and Jan Dumolyn. 2007. A short manual to the art of prosopography. In *Prosopography approaches and applications. A handbook*, pages 35–70. Unit for Prosopographical Research (Linacre College).

Christopher Warren, Daniel Shore, Jessica Otis, Lawrence Wang, Mike Finegold, and Cosma Shalizi. 2016. Six degrees of Francis Bacon: A statistical method for reconstructing large historical social networks. *Digital Humanities Quarterly*, 10(3).

Jun Zhao, Christian Bizer, A Gil, Paolo Missier, and Satya Sahoo. 2010. Provenance requirements for the next version of RDF. In *Proceedings of the W3C Workshop – RDF Next Steps*. W3C. `https://www.w3.org/2009/12/rdf-ws/papers/ws08`.

Publication VII

Mikko Koho, Erkki Heino, and Eero Hyvönen. SPARQL Faceter—Client-side Faceted Search Based on SPARQL. In *Joint Proceedings of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop co-located with the 13th Extended Semantic Web Conference ESWC 2016, Heraklion, Crete, Greece, May 30, 2016*, Raphaël Troncy, Ruben Verborgh, Lyndon Nixon, Thomas Kurz, Kai Schlegel, and Miel Vander Sande (editors), CEUR Workshop Proceedings, volume 1615, ISSN 16130073, online CEUR-WS.org/Vol-1615/semdevPaper5.pdf, May 2016.

SPARQL Faceter—Client-side Faceted Search Based on SPARQL

Mikko Koho, Erkki Heino, and Eero Hyvönen

Semantic Computing Research Group (SeCo),
Aalto University, Department of Computer Science, Finland
first.last@aalto.fi, http://seco.cs.aalto.fi/

Abstract. The faceted search paradigm is widely used in web applications, and there are various tools available for implementing it on the server side. In contrast, this paper presents an HTML based component tool on the client side that can be plugged on virtually any public SPARQL endpoint on the web, using *only* SPARQL API for data retrieval. To test and demonstrate the idea and the tool, application of the tool in a large in-use semantic portal is presented.

1 Introduction

Faceted search [2, 12], known also as view-based search [9] and dynamic hierarchies [10], is based on indexing data items along orthogonal category hierarchies, i.e., facets (e.g., places, times, document types etc.). In searching, the user selects in free order categories on facets, and the data items included in the selected categories are considered search results. After each selection, a count is computed for each category showing the number of results, if the user next makes that selection. In this way, search is guided by avoiding annoying "no hits" results. The idea of faceted search is especially useful on the Semantic Web where hierarhical ontologies used for data annotation provide a natural basis for facets, and reasoning can be used for mapping heterogeneous data to facets [4].

Faceted search can be implemented with popular server-side solutions, such as Solr[1], Sphinx[2], and ElasticSearch[3]. However there is a lack of light-weight client-side faceted search tools or components that search data directly from a SPARQL endpoint. Such a tool would be very useful, because it could be used on virtually any open SPARQL endpoint on the web without any need for server side programming and access rights. This paper presents a web component for implementing faceted search applications in a browser, based only on a standard SPARQL API.

[1] http://lucene.apache.org/solr/

[2] http://sphinxsearch.com/blog/2013/06/21/faceted-search-with-sphinx/

[3] https://www.elastic.co/

2 The Tool: SPARQL Faceter

Design Requirements The requirements for our tool, *SPARQL Faceter*, are based on our earlier experience on tediously implementing server side faceted search over and over again in different applications, as well as on the generic requirements of the search paradigm [12, 3]: there is a need for a lightweight browser-based faceted search engine tool that is easy to use and adapt for different RDF datasets published in SPARQL data services on the web. A particular demand of our own Linked Data Finland data service[4] [5] is to support external users of the data in application development, and here the notion of the Rich Internet Application, where the data service is completely detached using standard SPARQL API was deemed very useful. The tool should therefore fulfill the following requirements:

1. **Data read from a SPARQL endpoint.** It is a common practice to publish RDF datasets as SPARQL services, and build applications that use the SPARQL endpoint to query for information. The data can reside anywhere on the web and be managed by anybody on the server side. No extra effort is required for managing the data from the application developer side.
2. **Easily adaptable to different datasets.** Different datasets have different data models and the application should be easy to configure for these.
3. **Easy integration to web pages.** A developer should be able to integrate a faceted search on some web page and modify the appearance of the faceted search freely.
4. **Fluent user experience.** When a user makes selections of the facets, the application should show results in a reasonable time. The UI should be easy to use and should support the exploratory analysis of datasets.
5. **Support for hierarchical facets.** The tool should be able to handle hierarchical concepts in the facets, displaying the hierarchy in the facet selections. When a user selects a category upper in a hierarchy, the results should include all results for its' sub categories.
6. **Easy to maintain.** The tool is planned to be actively used, and thus should be easy to maintain and develop further.

In short, the goal was to create a tool that would be easy to use out-of-the-box, and require minimal configuration, yet provide enough configuration options to be useful in different contexts.

Design AngularJS[5] was chosen as the implementation framework for our tool. *SPARQL Faceter* is developed as open source with source code available on GitHub.

SPARQL Faceter consists of two distinct components. One is the main *Semantic Faceted Search* component[6], which contains the basic functionalities of

[4] http://ldf.fi

[5] https://angularjs.org/

[6] https://github.com/SemanticComputing/angular-semantic-faceted-search

the tool, and is dependent on the second component: an AngularJS service for querying SPARQL endpoints with paging and object mapping[7].

The *Semantic Faceted Search* component is comprised of different services, and most importantly, the facet selector directive. This directive is the main component of the tool, and provides the user interface for using the faceted search.

The facet selector directive is independent of any results the facet selections might imply: it's role is to keep track of what values are selected in what facets, and to display available selections to the user including the number of results each selection implies. The tool communicates the facet selections to whatever application is using it by calling a callback function whenever the selections change. Thus, *SPARQL Faceter* places no restrictions on what can be done with the user's selections. The tool does, however, provide services for integrating the facet selections to SPARQL queries, handling SPARQL results, and updating the URL based on current selections.

The *Semantic Faceted Search* component works by building a single SPARQL query for the facets it controls, and updating the query each time the user makes a selection. This has the advantage of keeping the states of the facets synchronized at all times. The downside is that with very many facets querying can become slow, and the user experience may suffer as the facets are unusable while their states are being queried. In order to make it feasible to have many facets available, facets can be enabled and disabled: only enabled facets are included in the query. Whether or not a facet is initially enabled is configurable. The tool shows a spinner on the components when querying for data, to convey to the user that the information in the elements is being updated.

For enabled facets, the tool shows all possible values for the facets' property, and the amount of instances that the selection results in. The amount of results is calculated based on all the current selections in all of the facets. Values that would lead to no results are omitted. In addition to the list of possible values, there is also a text input for searching the facet values.

The tool also supports hierarchical facets, in which selecting a category upper in the hierarchy also shows results for all the categories that are below it in the hierarchy. The hierarchy specification is not limited to using a predefined property or vocabulary, but is based on configuring a property path that captures the hierarchy, and giving the top categories explicitly. The facet element displays a two-level hierarchy of the categories. For categories below the current category level, the hierarchy is flattened and all the sub categories are listed below the upper category. The category level is initially on the topmost categories, and selecting a value changes category level to the level of the selected value.

The proposed way to use the facet selections, as returned by the tool, is to build another SPARQL query for retrieving the results according to the user's selections. By using the tool's services, the results can then be mapped into JavaScript objects with pagination and client-side caching of the received results.

[7] https://github.com/SemanticComputing/angular-paging-sparql-service

Installation and Configuration At its simplest, the only configuration options needed by *SPARQL Faceter* are the URL of a SPARQL endpoint, a callback function to be called with the facet selections when they change, the RDF class URI of the resources which are the target of the search, and the facet configurations. A basic facet can be configured with just the URI of the property and a name for the facet to be displayed to the user. The callback function defines what happens when the facet selections are updated, which typically is to update a results display according to user selections.

Other types of facets require some additional configuration options, but the amount of configuration needed for any facet type has been kept to a minimum. The other facet types include a keyword search facet, a time-span facet, and a hierarchical facet. More details about the configuration is given in the repository of the *Semantic Faceted Search* component.

As for the results, the user of the tool is expected to build a query for retrieving results based on the facet selections. This makes the tool extremely flexible to different data models, but requires that the user of the tool is familiar with SPARQL syntax.

3 Applications

We tested and evaluated *SPARQL Faceter* on the WarSampo portal [6]. The portal[8] consists of different application perspectives built on top of interlinked Finnish Second World War data published on a SPARQL endpoint. The SPARQL endpoint is provided by an Apache Fuseki Server[9]. The applications are tested to work with all major browsers (Chrome, Firefox and Internet Explorer).

Use case 1: Death Records Perspective The WarSampo death records perspective[10] uses the application to provide an interface for searching and exploring the death records of Finnish WW2 casualties [7]. The dataset consists of about 95,000 death records, and almost 2.4 million RDF triples.

Fig. 1 shows a screenshot of the faceted search of death records using *SPARQL Faceter*. The application interface contains 12 facets and a table-like view of the results. The facets include a keyword search facet for the persons' names, a time-span facet for the time of death and a hierarchical military rank facet, and 9 basic facets of the annotated properties of the death records. The source code of the application is available online[11].

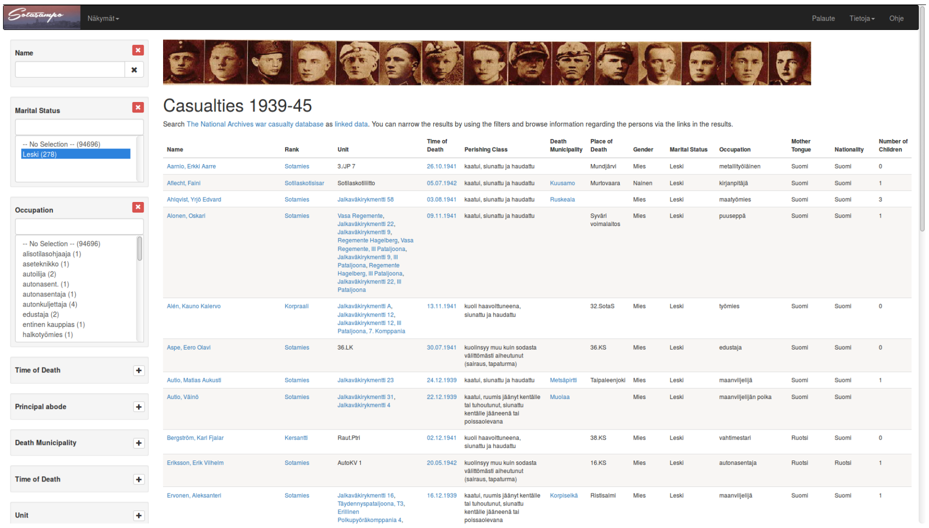
A dataset this large provides a benchmark for the performance of the *SPARQL Faceter*, as SPARQL queries and also data handling on the client-side could take long enough to deteriorate the user experience. Quite much effort had to be put on optimizing the SPARQL queries for speed, while also making sure that no

[8] http://www.sotasampo.fi/

[9] https://jena.apache.org/documentation/fuseki2/

[10] http://www.sotasampo.fi/en/casualties/

[11] https://github.com/SemanticComputing/WarSampo-death-records



Name

Marital Status
-- No Selection -- (94696)
Leski (278)

Occupation
-- No Selection -- (94696)
alisotilasohjaaja (1)
aseteknikko (1)
autoilija (2)
autonasent. (1)
autonasentaja (1)
autonkuljettaja (4)
edustaja (2)
entinen kauppias (1)
halkotyömies (1)

Time of Death

Principal abode

Death Municipality

Time of Death

Unit

Casualties 1939-45

Search The National Archives war casualty database as linked data. You can narrow the results by using the filters and browse information regarding the persons via the links in the results.

Name	Rank	Unit	Time of Death	Perishing Class	Death Municipality	Place of Death	Gender	Marital Status	Occupation	Mother Tongue	Nationality	Number of Children
Aarnio, Erkki Aarre	Sotamies	3./JP 7	26.10.1941	kaatui, siunattu ja haudattu		Mundjärvi	Mies	Leski	metallityöläinen	Suomi	Suomi	0
Aflecht, Faini	Sotilaskotisisar	Sotilaskotiliitto	05.07.1942	kaatui, siunattu ja haudattu	Kuusamo	Murtovaara	Nainen	Leski	kirjanpitäjä	Suomi	Suomi	1
Ahlqvist, Yrjö Edvard	Sotamies	Jalkaväkirykmentti 58	03.08.1941	kaatui, siunattu ja haudattu	Ruskeala		Mies	Leski	maatyömies	Suomi	Suomi	3
Alonen, Oskari	Sotamies	Vasa Regemente, Jalkaväkirykmentti 22, Jalkaväkirykmentti 9, Regemente Hagelberg, Vasa Regemente, III Pataljoona, Jalkaväkirykmentti 9, III Pataljoona, Regemente Hagelberg, III Pataljoona, Jalkaväkirykmentti 22, III Pataljoona	09.11.1941	kaatui, siunattu ja haudattu		Syväri voimalaitos	Mies	Leski	puuseppä	Suomi	Suomi	1
Alén, Kauno Kalervo	Korpraali	Jalkaväkirykmentti A, Jalkaväkirykmentti 12, Jalkaväkirykmentti 12, III Pataljoona, 7. Komppania	13.11.1941	kuoli haavoittuneena, siunattu ja haudattu		32.SotaS	Mies	Leski	työmies	Suomi	Suomi	0
Aspe, Eero Olavi	Sotamies	36.LK	30.07.1941	kuolinsyy muu kuin sodasta välittömästi aiheutunut (sairaus, tapaturma)		36.KS	Mies	Leski	edustaja	Suomi	Suomi	0
Autio, Matias Aukusti	Sotamies	Jalkaväkirykmentti 23	24.12.1939	kaatui, siunattu ja haudattu	Metsäpirtti	Taipaleenjoki	Mies	Leski	maanviljelijä	Suomi	Suomi	1
Autio, Väinö	Sotamies	Jalkaväkirykmentti 31, Jalkaväkirykmentti 4	22.12.1939	kaatui, ruumis jäänyt kentälle tai tuhoutunut, siunattu kentälle jääneenä tai poissaolevana	Muolaa		Mies	Leski	maanviljelijän poika	Suomi	Suomi	
Bergström, Karl Fjalar	Kersantti	Raut.Ptri	02.12.1941	kuoli haavoittuneena, siunattu ja haudattu		38.KS	Mies	Leski	vahtimestari	Ruotsi	Suomi	0
Eriksson, Erik Vilhelm	Sotamies	AutoKV 1	20.05.1942	kuolinsyy muu kuin sodasta välittömästi aiheutunut (sairaus, tapaturma)		16.KS	Mies	Leski	autonasentaja	Ruotsi	Ruotsi	1
Ervonen, Aleksanteri	Sotamies	Jalkaväkirykmentti 16, Täydennyspataljoona, T3, Erillinen Polkupyöräkomppania 4,	16.12.1939	kaatui, ruumis jäänyt kentälle tai tuhoutunut, siunattu kentälle jääneenä tai poissaolevana	Korpiselkä	Ristisalmi	Mies	Leski	maanviljelijä	Suomi	Suomi	1

Fig. 1. The faceted search interface of death records [7] with a selected value in one facet.

unnecessary processing of the results happens on the client-side. Normally when searching and browsing the dataset, results are displayed in a few seconds after the user makes a selection. With some selections that result in a large result set, and if additionally many facets are enabled, the user may have to wait more than ten seconds to see the results and updated facets. Most of the time goes to waiting for results from the SPARQL endpoint.

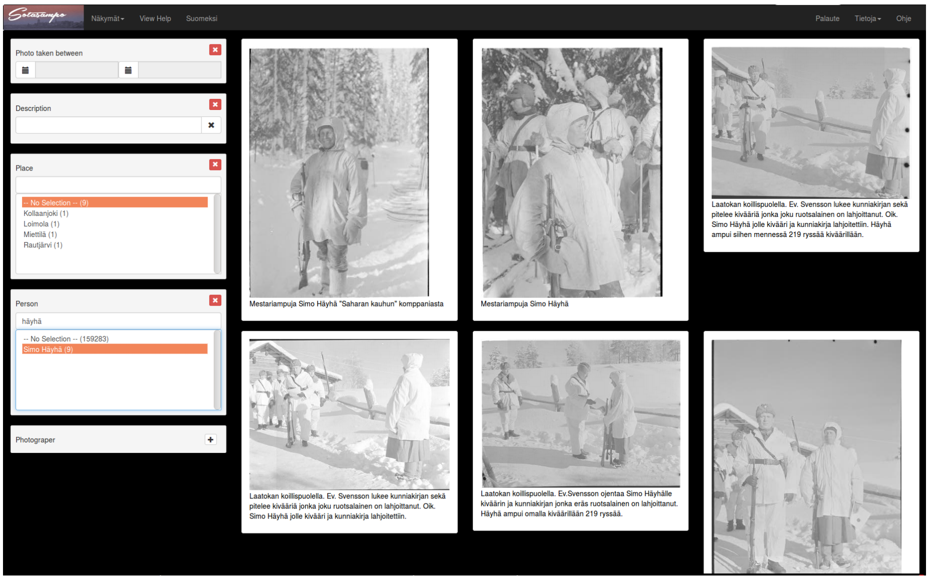
Use case 2: Photographs Perspective The WarSampo photographs perspective[12] uses the application to create a faceted search interface for wartime photographs. The dataset consists of about 159,000 photographs and a total of about 1.6 million triples. The interface contains only 5 facets, which ensures fast response times. The facets include a time-span facet for the date the photograph was taken, a keyword search for descriptions, and basic facets for related people and places. Fig. 2 shows a screenshot of the photograph perspective.

The perspective uses "infinite scrolling" to display the photographs: more results are retrieved as the user scrolls down for more photographs. Clicking on a photograph shows the user more information about the photograph, and provides hyperlinks to other WarSampo perspectives via linked entities.

4 Discussion

In this paper we have presented the *SPARQL Faceter* for creating client-side faceted search applications. The configuration of the facets has been kept as

[12] http://www.sotasampo.fi/en/photographs/



Place
-- No Selection -- (9)
Kollaanjoki (1)
Loimola (1)
Miettilä (1)
Rautjärvi (1)

Person: häyhä
-- No Selection -- (159283)
Simo Häyhä (9)

Mestariampuja Simo Häyhä "Saharan kauhun" komppaniasta

Mestariampuja Simo Häyhä

Laatokan koillispuolella. Ev. Svensson lukee kunniakirjan sekä pitelee kivääriä jonka joku ruotsalainen on lahjoittanut. Oik. Simo Häyhä jolle kivääri ja kunniakirja lahjoitettiin. Häyhä ampui siihen mennessä 219 ryssää kiväärillään.

Laatokan koillispuolella. Ev. Svensson lukee kunniakirjan sekä pitelee kivääriä jonka joku ruotsalainen on lahjoittanut. Oik. Simo Häyhä jolle kivääri ja kunniakirja lahjoitettiin.

Laatokan koillispuolella. Ev.Svensson ojentaa Simo Häyhälle kiväärin ja kunniakirjan jonka eräs ruotsalainen on lahjoittanut. Häyhä ampui omalla kiväärillään 219 ryssää.

Fig. 2. The faceted search interface of wartime photographs with a selected value in one facet.

simple as possible. *SPARQL Faceter* handles querying the SPARQL endpoint when needed, mapping the SPARQL results to JavaScript objects, and creating and updating the facet elements. To make use of the facet selection values, the developer using the tool has to retrieve results based on selected values, which usually includes writing SPARQL queries.

Results Section 2 presented the design requirements for a Semantic Web tool that provides client-side faceted search of RDF datasets published as SPARQL services. Here is a detailed view of how the *SPARQL Faceter* manages to satisfy these requirements:

1. **Data read from a SPARQL endpoint.** All data is retrieved via SPARQL, from an endpoint defined by the user.
2. **Easily adaptable to different datasets.** Configuration of the tool is required for every dataset. The configuration required for setting up facets is kept to a minimum.
3. **Easy integration to web pages.** A faceted search can be easily added to a web page by using the developed AngularJS components. The facet selections have a default look that can be customized using CSS.
4. **Fluent user experience.** For datasets that are distinctly smaller than the ones used in the use cases, the tool is almost certainly fast enough for good user experience. For the faceted search applications in the use cases, some user selections can be slow, but common usage gives results based on

user selections in a few seconds. For larger datasets, the amount of facets may have to be restricted, depending on the processing capabilities of the SPARQL server.

5. **Support for hierarchical facets.** Two-level hierarchical facets are supported by the tool, and demonstrated in the first use case. The dataset can contain more than two levels, but the hierarchical facet flattens the hierarchy to two levels for display in the select element.
6. **Easy to maintain.** The separation of concerns design principle is used in development to simplify code structure and maintenance. Automated tests are used to help maintain good code quality.

Our application is able to satisfy the given requirements, and thus provides a usable tool for faceted search of RDF datasets.

Related Work *Jassa* [11] is a JavaScript suite for SPARQL access, which among other things provides support for client-side faceted search. For creating simple faceted search applications Jassa is, however, a much larger and more complex solution than *SPARQL Faceter*.

Sparklis [1] is an application for exploring and querying data from a SPARQL endpoint using a natural language interface[13], which also implements a faceted search. *Sparklis* focuses on data exploration and requires no prior knowledge of the data model of the target endpoint, whereas *SPARQL Faceter* aims to support development of applications tailored for a specific dataset.

Oren et al. [8] have developed a prototype for faceted navigation of arbitrary RDF data with automatic facet selection, as a server-side solution.

Future Work Plans for future development include:

1. **Semi-automatic adaptation to new datasets.** The tool could create a basic configuration for a new SPARQL endpoint, possibly based on only a class selection by the user.
2. **Customizing facet appearance.** To support customizing the appearance of the facets further , we are planning a configuration option for the user to define their own template for the facets.

Acknowledgements Kasper Apajalahti and Jouni Tuominen contributed to an earlier version of our tool. This work was supported by the Linked Open Data Science project[14] funded by the Ministry of Education and Culture in Finland.

References

1. Ferré, S.: Expressive and scalable query-based faceted search over sparql endpoints. In: The Semantic Web–ISWC 2014, pp. 438–453. Springer (2014)

[13] http://www.irisa.fr/LIS/ferre/sparklis/

[14] http:/seco.cs.aalto.fi/projects/lodsci/

2. Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., Lee, K.P.: Finding the flow in web site search. CACM 45(9), 42–49 (2002)
3. Hearst, M.: Design recommendations for hierarchical faceted search interfaces. In: ACM SIGIR workshop on faceted search. pp. 1–5. Seattle, WA (2006)
4. Hyvönen, E., Saarela, S., Viljanen, K.: Application of ontology-based techniques to view-based semantic search and browsing. In: The semantic web: research and applications. First European Semantic Web Symposium (ESWS 2004). pp. 92–106. Springer-Verlag (2004)
5. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers. pp. 226–230. Springer-Verlag (May 2014)
6. 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). Springer-Verlag (May 2016)
7. 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: Proceedings of the 1st Workshop on Humanities in the Semantic Web - WHiSe (at ESWC 2016) (May 2016)
8. Oren, E., Delbru, R., Decker, S.: Extending faceted navigation for RDF data. In: The Semantic Web—ISWC 2006, pp. 559–572. Springer-Verlag (2006)
9. Pollitt, A.S.: The key role of classification and indexing in view-based searching. Tech. rep., University of Huddersfield, UK (1998), http://www.ifla.org/IV/ifla63/63polst.pdf
10. Sacco, G.M.: Dynamic taxonomies: guided interactive diagnostic assistance. In: Wickramasinghe, N. (ed.) Encyclopedia of Healthcare Information Systems. Idea Group (2005)
11. Stadler, C., Westphal, P., Lehmann, J.: Jassa: a JavaScript suite for SPARQL-based faceted search. In: Proceedings of the 2014 International Conference on Developers-Volume 1268. pp. 31–36. CEUR-WS. org (2014)
12. Tunkelang, D.: Faceted search, Synthesis lectures on information concepts, retrieval, and services, vol. 1. Morgan & Claypool Publishers (2009)

Publication VIII

Mikko Koho, Eero Hyvönen, Erkki Heino, Jouni Tuominen, Petri Leskinen, and Eetu Mäkelä. Linked Death — Representing, Publishing, and Using Second World War Death Records as Linked Open Data. In *The Semantic Web: ESWC 2017 Satellite Events: ESWC 2017 Satellite Events, Portorož, Slovenia, May 28 – June 1, 2017, Revised Selected Papers*, Eva Blomqvist, Katja Hose, Heiko Paulheim, Agnieszka Ławrynowicz, Fabio Ciravegna, and Olaf Hartig (editors), Lecture Notes in Computer Science, volume 10577, pages 369–383, ISBN 9783319704067, Springer, Cham, May–June 2017.

Linked Death—Representing, Publishing, and Using Second World War Death Records as Linked Open Data

Mikko Koho, Eero Hyvönen, Erkki Heino, Jouni Tuominen, Petri Leskinen, and Eetu Mäkelä

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

Abstract. War history of the Second World War (WW2), humankind's largest disaster, is of great interest to both laymen and researchers. Most of us have ancestors and relatives who participated in the war, and in the worst case got killed. Researchers are eager to find out what actually happened then, and even more importantly why, so that future wars could perhaps be prevented. The darkest data of war history are casualty records—from such data we could perhaps learn most about the war. This paper presents a model and system for representing death records as linked data, so that 1) citizens could find out more easily what happened to their relatives during WW2 and 2) digital humanities (DH) researchers could (re)use the data easily for research.

1 Introduction

Lots of information about the WW2 is available on the web[1]. However, this information is typically meant for human consumption only. The underlying *data* is not available in machine-readable, i.e., "semantic" form for Digital Humanities research and use [5,3] and for end-user applications to utilize. 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 hopefully promotes peace in the future [8].

For the case of the First World War, the situation has started to change, with several projects publishing linked data on the web, such as Europeana Collections 1914–1918[2], 1914–1918 Online[3], WW1 Discovery[4], Out of the Trenches[5], CENDARI[6], Muninn[7], and WW1LOD [12]. A few works have used the linked

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

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

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

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

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

[6] `http://www.cendari.eu/research/first-world-war-studies/`

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

data approach to WW2 data as well, such as [2,1], the Open Memory Project[8], and WarSampo [8].

This paper discusses the publication and use of casualty (death) records as linked data, as one part of the larger WarSampo system. Here, a dataset of some 95,000 deaths in military action in the Finnish army is concerned. We first present the data, its modeling, the Linked Open Data (LOD) service, and interlinking the data with other WarSampo datasets. After this, two use case applications are presented: 1) analyzing the data for digital humanities research and 2) reassembling the biographical war history of individual soldiers and military units. The latter use case serves, e.g., laymen in trying to figure out what happened to their relatives in WW2. The WarSampo system[9] was published on Nov 27, 2015 and has had tens of thousands of end users indicating a large public interest in such applications.

2 Dataset, Data Model, and Data Service

Information about all known Finnish casualties of WW2 has been gathered in a relational database at the National Archives. This database contains 94,696 records of people that fought on the Finnish side, and died in 1939–1945 in the Winter War, the Continuation War, or in the Lapland War, or died of injuries obtained in those wars.

For use in the WarSampo project, the casualty database was first converted to CSV format, which was then converted to RDF format. Because the objective was to develop interactive applications directly on top of the large RDF dataset, it was important to keep the amount of RDF triples as low as possible without losing information and still linking the death records to ontological concepts. Thus, a simple data model was created for representing the data as linked data.

The data model is based on the CIDOC Conceptual Reference Model (CRM) vocabulary, which is designed for information exchange and integration of various cultural heritage information [4]. Each death record is represented as an instance of the Document class (`crm:E31_Document`) of CIDOC CRM.

A metadata schema was created that defines the properties used to describe each casualty with the information from the original database. The schema consists of OWL properties which have `crm:E31_Document` as the domain. A list of the properties and their `rdfs:range` constraints are shown in Table 1. The namespace prefixes used in this paper are:

:	http://ldf.fi/schema/narc-menehtyneet1939-45/
crm:	http://www.cidoc-crm.org/cidoc-crm/
skos:	http://www.w3.org/2004/02/skos/core#
wat:	http://ldf.fi/warsa/actors/actor_types/
wrank:	http://ldf.fi/warsa/actors/ranks/

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

[9] Including a semantic portal in use at `http://sotasampo.fi` and the underlying LOD SPARQL service at `http://www.ldf.fi/dataset/warsa/`.

Table 1. Casualty metadata schema of all properties used for describing the death records.

Property description	Property name	Range
mother tongue	:aeidinkieli	:Aeidinkieli
occupation	:ammatti	xsd:string
principal abode	:asuinkunta	
first names	:etunimet	xsd:string
date of becoming wounded	:haavoittumisaika	xsd:date
municipality of becoming wounded	:haavoittumiskunta	
place of becoming wounded	:haavoittumispaikka	xsd:string
burial place	:hautapaikka	xsd:string
burial graveyard	:hautausmaa	:Hautausmaa
military unit	:joukko_osasto	xsd:string
military unit code	:joukko_osastokoodi	xsd:string
known military unit	:osasto	wat:MilitaryUnit
citizenship	:kansalaisuus	:Kansalaisuus
nationality at time of death	:kansallisuus	:Kansallisuus
date of becoming missing	:katoamisaika	xsd:date
municipality of becoming missing	:katoamiskunta	
place of becoming missing	:katoamispaikka	xsd:string
place of domicile	:kotikunta	
date of death	:kuolinaika	xsd:date
municipality of death	:kuolinkunta	
place of death	:kuolinpaikka	xsd:string
number of children	:lasten_lukumaeaerae	xsd:integer
perishing class	:menehtymisluokka	:Menehtymisluokka
marital status	:siviilisaeaety	:Siviilisaeaety
military rank	:sotilasarvo	wrank:Rank
last name	:sukunimi	xsd:string
gender	:sukupuoli	:Sukupuoli
municipality of birth	:synnyinkunta	
date of birth	:syntymaeaika	xsd:date
full name	skos:prefLabel	rdfs:Literal
WarSampo person instance	crm:P70_documents	crm:E21_Person

The default namespace corresponds to the casualty schema namespace. RDF Schema (RDFS), Web Ontology Language (OWL) and XML Schema namespaces are omitted.

In Table 1 there are a total of 31 properties that are used for describing the casualties. The properties are used only when there is a value for the property. Municipalities are currently linked to three distinct datasets, which is why their range is not defined. The place properties, which give a more specific place for the described events, are literals representing the place names of the original data. Original text representations of military units are also preserved and a new property `:osasto` is added for linking to WarSampo military units.

The Simple Knowledge Organization System (SKOS)[10] was used to define vocabularies to present the information found in the original database in RDF. The created SKOS vocabularies for describing the death records in the casualty dataset are listed in table 2.

Table 2. SKOS vocabularies for describing the death records.

Vocabulary	**Number of concepts**
citizenships	10
genders	3
graveyards	802
marital statuses	5
mother tongues	11
municipalities	632
nationalities	11
perishing classes	7

A graveyard vocabulary was created to describe graveyards around Finland, and is also linked to ontologies of Finnish municipalities. These municipalities include current municipalities as well as historical municipalities, as some graveyards are located outside current Finnish borders, and often only the historical municipality of the graveyard is known.

The dataset is published on the Linked Data Finland (LDF) [7] platform, where it is openly available[11] for use via a SPARQL endpoint, with the Creative Commons Attribution 4.0 license[12]. The SPARQL endpoint[13] serves all WarSampo data, and has distinct graphs for each separate dataset and a default graph which contains all data. A Fuseki[14] SPARQL Server is used for storing and serving the linked data. The used URIs are dereferenceable and provide information about resources for both human and machine users.

3 Interlinking with WarSampo datasets

The RDF dataset has been enriched by linking it to other parts of WarSampo like military ranks, military units, information about people found in other sources, and municipalities of wartime Finland.

A figure displaying the external linking of the death records is shown in Fig. 1. Each casualty is linked to other related WarSampo datasets and to a common

[10] `https://www.w3.org/2009/08/skos-reference/skos.html`

[11] `http://www.ldf.fi/dataset/narc-menehtyneet1939-45`

[12] `https://creativecommons.org/licenses/by/4.0/`

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

[14] `http://jena.apache.org/documentation/serving_data/`

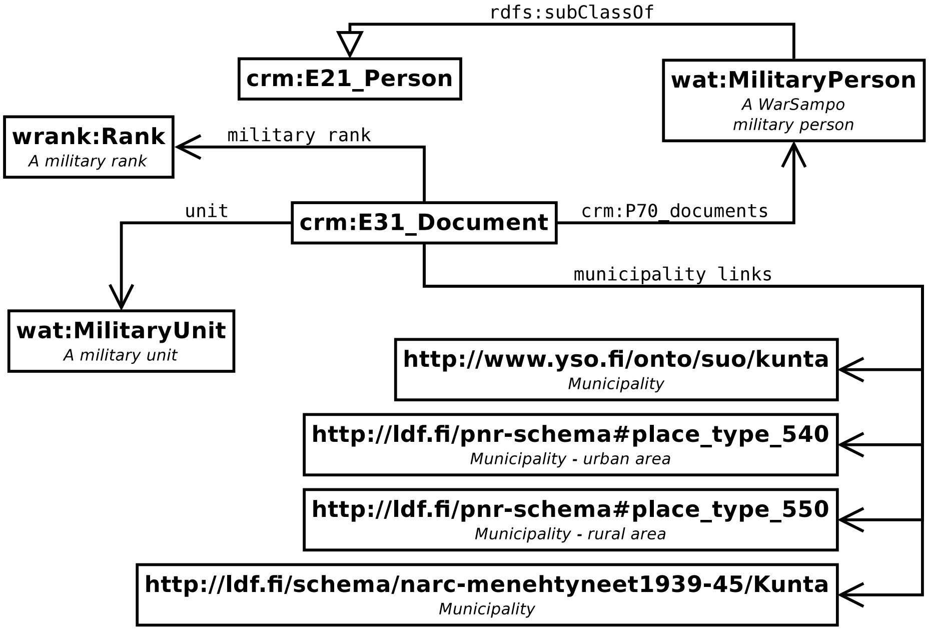


Fig. 1. External links from the death records, which are instances of `crm:E31_Document`. The properties are explained in Table 1. All WarSampo classes are aligned with the CIDOC CRM framework using `rdfs:subClassOf` relations.

WarSampo military rank ontology. The annotated military rank represents the rank of a person at the time of death.

Each death record is linked to a person instance (crm:E21_Person) of the WarSampo persons dataset via `crm:P70_documents`. So in our model, the death records are documents about the actual person. New information about people from other sources is not added directly to the casualty dataset, but to the WarSampo persons dataset, in order to maintain the integrity of the casualty dataset as a whole.

There are 3 municipality datasets, which include a dataset of contemporary Finnish municipalities, part of the Finnish Geographic Names Registry[15], and two datasets that complement each other and consist of historical Finnish municipalities. These are the historical wartime municipalities registry [6] of WarSampo, and a municipality ontology based on the casualty dataset. The place properties are currently not linked to places available [6] in the Karelian map names in Finland and Russia, and the Finnish Geographic Names Registry. This is because finding the places to link automatically would result in a low precision due to the abundance of distinct places with identical names.

The death records are programmatically linked with people found in other WarSampo sources, which are gathered in the WarSampo persons dataset. The

[15] `http://www.ldf.fi/dataset/pnr`

linking is implemented using the automatic annotation service ARPA [11], and a fuzzy logic algorithm to calculate a score for the similarity of two people based on each person's name, birth date, death date, and military rank. Source code for all automatic linking is available on GitHub[16].

For each death record, the annotation service first generates a set of candidates with similar names. The candidates are then scored and if the score is above a given threshold, the persons are expected to be the same and are linked. Similarity score is given for string similarity for first names and last name, if they are similar enough, to allow somewhat differing spellings of names. A score is given for matches in military rank, birth date and death date, and subtracted in case they are not matching. In scoring, any of the values are allowed to be missing. In case of multiple matches for a death record, only the best match is used.

A `crm:P70_documents` relation is then added to the death record that will be linked, that points to the matching person. After this new person instances are created to WarSampo persons dataset for each death record that were not linked to an existing person.

We were able to automatically link 118 death records to people gathered from other data sources in the WarSampo persons dataset. The amount is quite low because the person information from other sources currently contains mostly information about high ranking officers and people who survived the war. The found person links have been manually validated and the discovered links seem to be depicting the same people. Manual validation was also done to person pairs that were close to the score threshold but not linked, and these seem to either not depict the same persons or not have enough information to make an assumption either way. However, as the scoring is manually adjusted to work well for the current persons dataset, when new people from new sources are added to the persons dataset, the scoring may need readjusting.

Military units of casualties are also programmatically linked to military units described in the WarSampo army units dataset, which contains military unit information found in other sources. The linking is implemented using the ARPA service and is based on unit abbreviations found in the casualty dataset, which are matched against manually annotated unit abbreviations in the army units dataset. As the exact abbreviation formats vary somewhat in different sources, multiple different abbreviation formats are generated from the original one for use in the automatic linking. Some 66,700 death records were linked to WarSampo military units, so this accounts for 70% of all the casualties. Currently, military unit information in WarSampo is limited to units of the Winter War. Therefore not all casualties are linked to the military units of WarSampo.

Municipalities in the data are linked automatically based on the labels of the municipalities. As shown in Table 1, there are six properties that relate to municipalities for each death record. The automatic linking leads to 98% of all death records having at least one link to the known wartime municipalities, which is the primary municipality dataset of WarSampo.

[16] `https://github.com/SemanticComputing/Casualty-linking`

4 Use Case 1: Studying Death Records

This use case studies how the data could be used for prosopographical digital humanities research. We present the casualties perspective of the WarSampo portal, which is a tool for interactively analyzing the data in order to find patterns in groups of individuals.

The dataset graph consists of almost 2.4 million RDF triples. Presenting the data in an online service for users to search and browse is not straightforward due to the large size of the dataset. Furthermore, there are lots of links to related data in other WarSampo datasets (people, places, military units, etc.).

Faceted search provides effective support for interactive information-seeking in information systems [13]. A faceted search application was developed for searching and browsing the dataset. The application[17] is part of the WarSampo portal, and provides the casualties perspective as one of the portal's different perspectives. Faceted search is based on displaying categories for each facet, from which the user can select one, which then narrow down the result set to include only the results that match the user selections.

Fig. 2 shows a screenshot of the faceted search application in the casualties perspective. The data is laid out in a table-like view. Facets are presented on the left of the interface with string search support. The number of hits on each facet is calculated dynamically and shown to the user, and selections leading to an empty result set are hidden.

In Fig. 2, seven facets and the results are shown, where the user has selected "widow" in the marital status facet, focusing the search down to 278 killed widows that are presented in the table with links to further information.

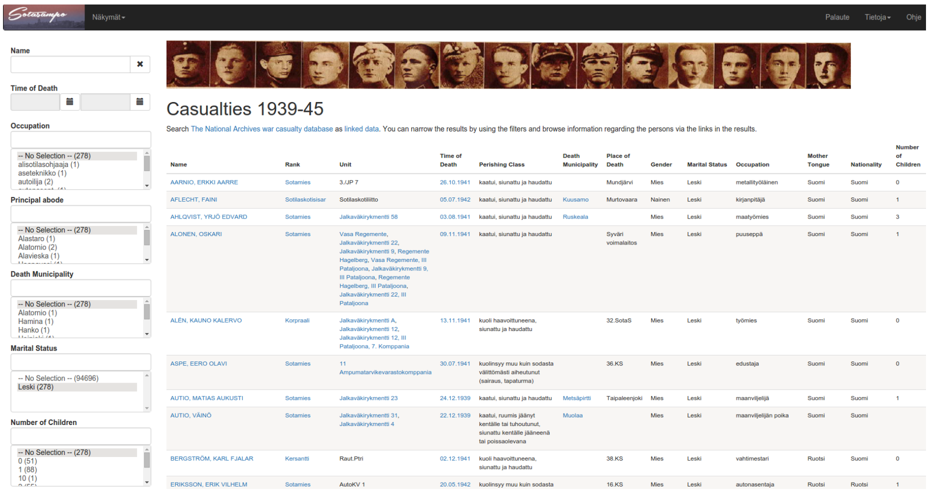
The faceted search is used not only for searching but also as a flexible tool for researching the underlying data. In Fig. 2, the hit counts immediately show distributions of the killed widows along the facet categories. For example, the facet "Number of children" shows that one of the deceased had 10 children and most often (in 88 cases) widows had one child. If we next select the category "one child" on its facet, we can see that two of the deceased are women and 86 are men in the gender facet.

The application is developed in JavaScript as a Rich Internet Application (RIA) on the client side, using the open SPARQL endpoint to fetch data according to user selections. The application is open source[18], and is based on our *SPARQL Faceter* tool [9], which is also open source[19]. When the user's selections of the facets change, an asynchronous SPARQL query is sent from the user's web browser to the SPARQL endpoint. The SPARQL endpoint returns results of the query to the user's browser, which does additional processing of the data before displaying the new results to the user. The system works well even with the large casualty dataset, because pagination is used to limit the amount of results that are queried and displayed at a time.

[17] `http://www.sotasampo.fi/casualties/`

[18] `https://github.com/SemanticComputing/WarSampo-death-records`

[19] `https://github.com/SemanticComputing/angular-semantic-faceted-search`



Name

Time of Death

Occupation

-- No Selection -- (278)
alisotilasohjaaja (1)
aseteknikko (1)
autoilija (2)

Principal abode

-- No Selection -- (278)
Alastaro (1)
Alatornio (2)
Alavieska (1)

Death Municipality

-- No Selection -- (278)
Alatornio (1)
Hamina (1)
Hanko (1)

Marital Status

-- No Selection -- (94696)
Leski (278)

Number of Children

-- No Selection -- (278)
0 (51)
1 (88)
10 (1)

Casualties 1939-45

Search The National Archives war casualty database as linked data. You can narrow the results by using the filters and browse information regarding the persons via the links in the results.

Name	Rank	Unit	Time of Death	Perishing Class	Death Municipality	Place of Death	Gender	Marital Status	Occupation	Mother Tongue	Nationality	Number of Children
AARNIO, ERKKI AARRE	Sotamies	3./JP 7	26.10.1941	kaatui, siunattu ja haudattu		Mundjärvi	Mies	Leski	metallityöläinen	Suomi	Suomi	0
AFLECHT, FAINI	Sotilaskotisisar	Sotilaskotiliitto	05.07.1942	kaatui, siunattu ja haudattu	Kuusamo	Murtovaara	Nainen	Leski	kirjanpitäjä	Suomi	Suomi	1
AHLQVIST, YRJÖ EDVARD	Sotamies	Jalkaväkirykmentti 58	03.08.1941	kaatui, siunattu ja haudattu	Ruskeala		Mies	Leski	maatyömies	Suomi	Suomi	3
ALONEN, OSKARI	Sotamies	Vasa Regemente, Jalkaväkirykmentti 22, Jalkaväkirykmentti 9, Regemente Hagelberg, Vasa Regemente, III Pataljoona, Jalkaväkirykmentti 9, III Pataljoona, Regemente Hagelberg, III Pataljoona, Jalkaväkirykmentti 22, III Pataljoona	09.11.1941	kaatui, siunattu ja haudattu		Syväri voimalaitos	Mies	Leski	puuseppä	Suomi	Suomi	1
ALÉN, KAUNO KALERVO	Korpraali	Jalkaväkirykmentti A, Jalkaväkirykmentti 12, Jalkaväkirykmentti 12, III Pataljoona, 7. Komppania	13.11.1941	kuoli haavoittuneena, siunattu ja haudattu		32.SotaS	Mies	Leski	työmies	Suomi	Suomi	0
ASPE, EERO OLAVI	Sotamies	11 Ampumatarvikevarastokomppania	30.07.1941	kuolinsyy muu kuin sodasta välittömästi aiheutunut (sairaus, tapaturma)		36.KS	Mies	Leski	edustaja	Suomi	Suomi	0
AUTIO, MATIAS AUKUSTI	Sotamies	Jalkaväkirykmentti 23	24.12.1939	kaatui, siunattu ja haudattu	Metsäpirtti	Taipaleenjoki	Mies	Leski	maanviljelijä	Suomi	Suomi	1
AUTIO, VÄINÖ	Sotamies	Jalkaväkirykmentti 31, Jalkaväkirykmentti 4	22.12.1939	kaatui, ruumis jäänyt kentälle tai tuhoutunut, siunattu kentälle jääneenä tai poissaolevana	Muolaa		Mies	Leski	maanviljelijän poika	Suomi	Suomi	
BERGSTRÖM, KARL FJALAR	Kersantti	Raut.Ptri	02.12.1941	kuoli haavoittuneena, siunattu ja haudattu		38.KS	Mies	Leski	vahtimestari	Ruotsi	Suomi	0
ERIKSSON, ERIK VILHELM	Sotamies	AutoKV 1	20.05.1942	kuolinsyy muu kuin sodasta		16.KS	Mies	Leski	autonasentaja	Ruotsi	Ruotsi	1

Fig. 2. The faceted search interface of death records with one selected facet. The left side contains the facets, displaying available categories and the amount of death records for each casualty. Death records matching the current facet selections are shown as a table.

The faceted search application uses the AngularJS framework, and is based on two distinct components that together provide the needed functionalities: the main faceted search component, that handles the user interface, and a SPARQL service that maps results to JavaScript objects.

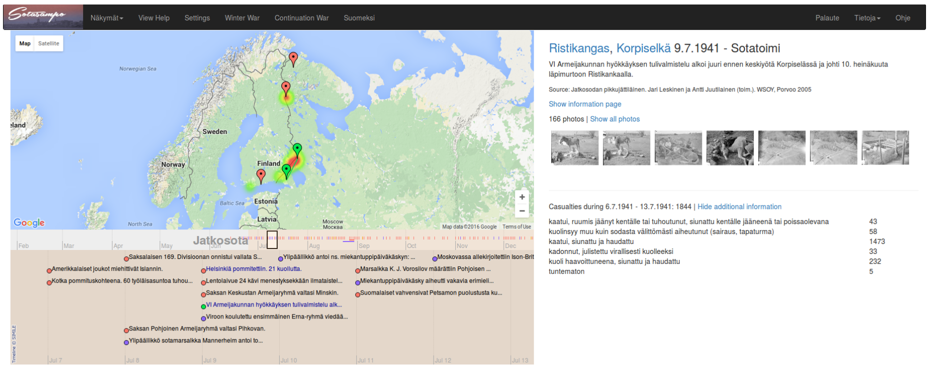
The application builds a single SPARQL query that retrieves the facet categories and the amount of results for each of them. Another query is built for the results display.

The shown facets are configured in the application and they directly use the properties of each death record instance. The categories shown in facets are values of properties of the death records, which may be resources selected from the corresponding ontologies, such as places, or plain literals.

The interface contains 12 facets, of which nine are basic facets that display the values of a property as categories. In addition to the basic facets, there are three facets with different functionalities. The text search facet is used for finding people directly by name: the user just enters a person's name or a part of it into the search box. The date of death facet has a date range selector to filter the results. The military rank facet is a hierarchical facet, making use of the hierarchical nature of the military ranks. The used military rank ontology contains two hierarchies, one based on the actual rank, and one for grouping ranks to, e.g., generals, officers, and enlisted ranks. Of these, the rank group hierarchy is used in the facet. Selecting a category upper in the hierarchy also shows results for all the categories that are below it in the hierarchy. The hierarchy is flattened to show only two distinct levels. The level of the current selection, or initially the top level, is shown on an upper level and the values that are lower in the hierarchy are displayed on a lower level below the corresponding upper level categories.

Most of the facets are disabled by default, and the user has to click a plus sign on the facet to activate it. Activating facets makes the interface respond more slowly to user selections, as data for each activated facet has to be queried from the SPARQL endpoint based on user selections to show the facet categories. Normally when a user searches or browses the dataset, the facet categories and the results display are updated within a few seconds after the user has made a selection from one of the facets. With selections that have a large result set, and if additionally many facets are enabled, the user may have to wait more than ten seconds.

Another perspective of the WarSampo portal that makes use of the casualty data is the event perspective. The perspective displays wartime events on a timeline and map, as seen in Fig. 3. The casualty data is visualized by a heat map layer on the map, showing an overview of where casualties occurred during different time frames, and also which events happened nearby. The application also displays statistics regarding the casualties during the selected time frame: the total amount of casualties, and the amount per perishing class. People mentioned in the event descriptions are linked to the WarSampo persons dataset, and through this link to the casualties dataset for people that have died in the wars. The application provides hyperlinks to the linked entities shown in their corresponding perspectives. These other perspectives include applications for exploring places, photographs, military units, and magazine articles of the WarSampo system.



Saksalaisen 169. Divisioonan onnistui vallata S...

Ylipäällikkö antoi ns. miekantuppipäiväkäskyn: ...

Moskovassa allekirjoitettiin Ison-Brit

Amerikkalaiset joukot miehittivät Islannin.

Helsinkiä pommitettiin. 21 kuollutta.

Marsalkka K. J. Vorosilov määrättiin Pohjoisen ...

Kotka pommituskohteena. 60 työläisasuntoa tuhou...

Lentolaivue 24 kävi menestyksekkään ilmataistel...

Miekantuppipäiväkäsky aiheutti vakavia erimieli...

Saksan Keskustan Armeijaryhmä valtasi Minskin.

Suomalaiset vahvensivat Petsamon puolustusta ku...

VI Armeijakunnan hyökkäyksen tulivalmistelu alk...

Viroon koulutettu ensimmäinen Erna-ryhmä viedää...

Saksan Pohjoinen Armeijaryhmä valtasi Pihkovan.

Ylipäällikkö sotamarsalkka Mannerheim antoi to...

Ristikangas, Korpiselkä 9.7.1941 - Sotatoimi

VI Armeijakunnan hyökkäyksen tulivalmistelu alkoi juuri ennen keskiyötä Korpiselässä ja johti 10. heinäkuuta läpimurtoon Ristikankaalla.

Source: Jatkosodan pikkujättiläinen. Jari Leskinen ja Antti Juutilainen (toim.). WSOY, Porvoo 2005

Show information page

166 photos | Show all photos

Casualties during 6.7.1941 - 13.7.1941: 1844 | Hide additional information

kaatui, ruumis jäänyt kentälle tai tuhoutunut, siunattu kentälle jääneenä tai poissaolevana	43
kuolinsyy muu kuin sodasta välittömästi aiheutunut (sairaus, tapaturma)	58
kaatui, siunattu ja haudattu	1473
kadonnut, julistettu virallisesti kuolleeksi	33
kuoli haavoittuneena, siunattu ja haudattu	232
tuntematon	5

Fig. 3. The WarSampo event perspective with casualties of an 8 day time period visualized as a heat map on top of Google Maps, and casualty statistics in the bottom right corner.

5 Use Case 2: Reassembling Soldier Biographies and Military Unit Histories

This use case studies how we can reassemble soldier biographies and military unit histories based on the information content available in the casualties dataset and linked information in other WarSampo datasets. This use case serves citizens and researchers who are interested in finding information about a person's involvement in the war.

Linking the death records to information about the same people in other sources, events, military units, war diaries, photographs and wartime places provides new information about their activities, involvements in war events, whereabouts and movements during the war. By linking all these pieces of information together, we are able to construct partial biographies of individual soldiers, and the movements and actions of their military units. This allows an individual who is interested in investigating the biography of a relative who took part in the war to look at where the person probably fought, with whom, and when, and in what events his military unit participated. Also, the interlinked dataset, together with applications to effectively use it, provides digital humanities researchers with new perspectives to study the casualties, that would not be possible with a non-linked dataset.

Fig. 4 shows a histogram of the amount of casualties per day of a single military unit, the Second Battalion of the 38th Infantry Regiment, and all of its subunits, which consist of 4 companies. The battalion existed during the Winter War, which was fought from 30 November 1939 to 13 March 1940. The time span in the figure covers the time from first casualty to the last, with the exception of one death that occurred much later in 1940, supposedly due to injuries obtained in the war.

Demonstrating the value of linking additional data to the death records, we have information of 19 events that are linked to the military unit and its subdivisions. They seem to explain quite well the casualties during the Winter War, as high peaks in casualties mostly occur when the unit is engaged in an

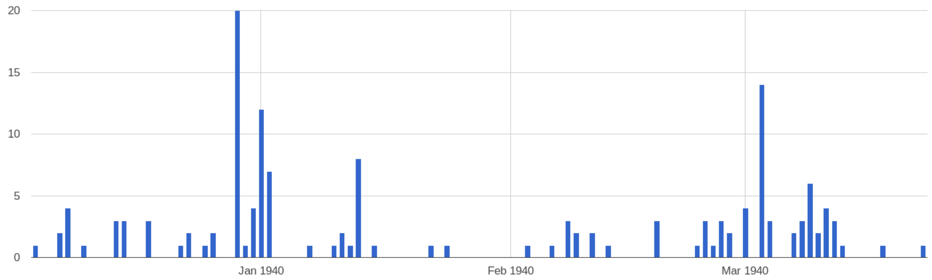


Fig. 4. A histogram visualization which shows casualties of the the Second Battalion of the 38th Infantry Regiment during the Winter War.

assault. However, the second highest peak occurs during a long defensive battle just before the end of the Winter War.

Events of the whole 38th Infantry Regiment during the Winter War are quite well covered in WarSampo. The regiment's second battalion and its subdivisions are known to have participated in the following events:

1. Defensive battle at Lavajärvi, 1939-12-06 – 1939-12-08
2. Battle at Lavajärvi, 1939-12-07 – 1939-12-08
3. Stalling battle at Lavajärvi-Lemetti, 1939-12-08 – 1939-12-10
4. Assault on Karjamökki, 1939-12-14 – 1939-12-14
5. Defense of Syskyjärvi sector, 1939-12-15 – 1939-12-28
6. Assault on Ruhtinaanmäki, 1939-12-29 – 1940-01-03
7. Assault on western Lemetti, 1940-01-06 – 1940-01-07
8. Assault battles at Repomäki, 1940-01-08 – 1940-01-12
9. Assault on Ruunaviita, 1940-01-13 – 1940-01-15
10. Battle at Koivuselkä, 1940-01-17 – 1940-02-06
11. Assault on Kehnovaara, 1940-01-18 – 1940-01-18
12. Assault on hill 63 and its defense, 1940-01-21 – 1940-01-24
13. Occupation of Pukitsanmäki and its defense, 1940-01-23 – 1940-01-26
14. Assault on Pujaski-Borisoff, 1940-01-25 – 1940-01-25
15. Destruction of Soviet elite ski unit at south of western Lemetti, 1940-02-06 – 1940-02-06
16. Capturing an encirclement northeast of Nietjärvi and destruction of a Soviet elite ski battalion, 1940-02-07 – 1940-02-09
17. Capturing 3 encirclements at Konnunkylä (Pujaski, Ahola and between railroad and road), 1940-02-18 – 1940-02-19
18. Capturing an encirclement south of point 26 (about 200 metres east of Koivusilta), 1940-02-21 – 1940-02-21
19. Defensive battle south of Nietjärvi, 1940-02-24 – 1940-03-11

The histogram is created by reading data directly from the WarSampo SPARQL endpoint and visualizing it with YASGUI[20] online SPARQL tool. The SPARQL query for retrieving the casualties for this military unit and its subdivisions is the following:

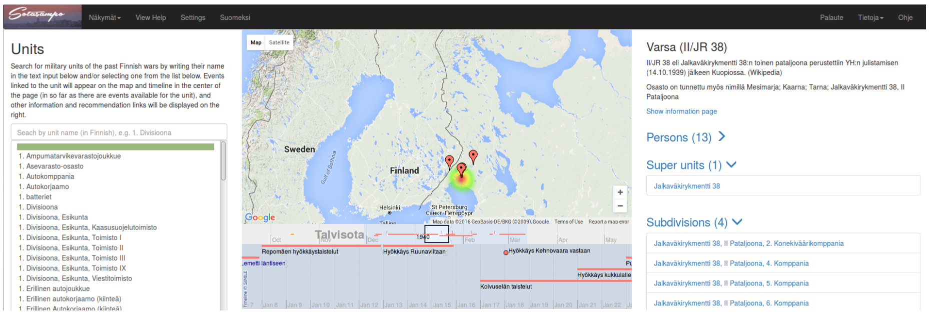
```
PREFIX atypes: <http://ldf.fi/warsa/actors/actor_types/>
PREFIX crm: <http://www.cidoc-crm.org/cidoc-crm/>
PREFIX casualties: <http://ldf.fi/schema/narc-menehtyneet1939-45/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?date (count(?cas) as ?casualties)
WHERE {
  { SELECT ?subunit
    WHERE {
      VALUES ?unit { <http://ldf.fi/warsa/actors/actor_972> } .
      ?unit (^crm:P144_joined_with/crm:P143_joined)+ ?subunit .
      ?subunit a atypes:MilitaryUnit .
    }
  } UNION {
   VALUES ?subunit { <http://ldf.fi/warsa/actors/actor_972> } .
```

[20] http://yasgui.org

```
  }
  ?cas casualties:osasto ?subunit .
  ?cas casualties:kuolinaika ?date .
  FILTER(?date < "1940-06-01"^^xsd:date)
} GROUP BY ?date ORDER BY ?date
```

All of the information about the Second Battalion of the 38th Infantry Regiment in WarSampo are available through the units perspective of WarSampo[21]. The units perspective visualizes the troop actions both on a map and a timeline, and shows the casualties of the unit as a heat map in the same fashion as in the event perspective. A screenshot of the perspective is shown in Fig. 5.

For even deeper understanding of the history, links to digitized images of the war diaries of the army units are provided, containing rich primary source descriptions of the events. By following the municipality links to the places perspective of WarSampo one can, e.g., study what kind of war events took place in the person's birth place, see photographs taken at specific locations the troops were located in, or read magazine articles depicting wartime events that took place in some specific place.



Units

Search for military units of the past Finnish wars by writing their name in the text input below and/or selecting one from the list below. Events linked to the unit will appear on the map and timeline in the center of the page (in so far as there are events available for the unit), and other information and recommendation links will be displayed on the right.

Seach by unit name (in Finnish), e.g. 1. Divisioona

1. Ampumatarvikevarastojoukkue
1. Asevarasto-osasto
1. Autokomppania
1. Autokorjaamo
1. batteriet
1. Divisioona
1. Divisioona, Esikunta
1. Divisioona, Esikunta, Kaasusuojelutoimisto
1. Divisioona, Esikunta, Toimisto I
1. Divisioona, Esikunta, Toimisto II
1. Divisioona, Esikunta, Toimisto III
1. Divisioona, Esikunta, Toimisto IX
1. Divisioona, Esikunta, Viestitoimisto
1. Erillinen autojoukkue
1. Erillinen autokorjaamo (kiinteä)
1. Erillinen Autokorjaamo (kiinteä)

Varsa (II/JR 38)

II/JR 38 eli Jalkaväkirykmentti 38:n toinen pataljoona perustettiin YH:n julistamisen (14.10.1939) jälkeen Kuopiossa. (Wikipedia)

Osasto on tunnettu myös nimillä Mesimarja; Kaarna; Tarna; Jalkaväkirykmentti 38, II Pataljoona

Show information page

Persons (13)

Super units (1)

Jalkaväkirykmentti 38

Subdivisions (4)

Jalkaväkirykmentti 38, II Pataljoona, 2. Konekiväärikomppania
Jalkaväkirykmentti 38, II Pataljoona, 4. Komppania
Jalkaväkirykmentti 38, II Pataljoona, 5. Komppania
Jalkaväkirykmentti 38, II Pataljoona, 6. Komppania

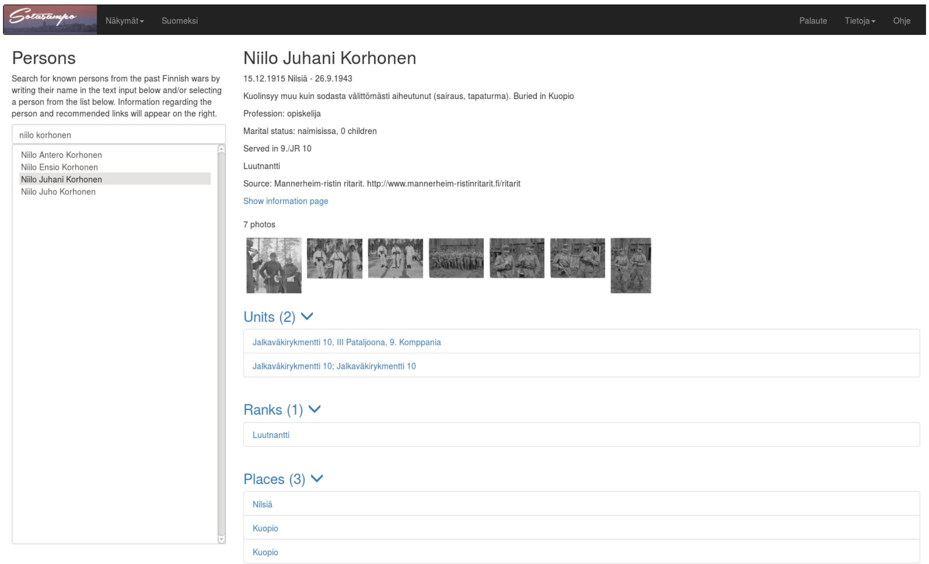
Fig. 5. WarSampo military units perspective displaying information about the Second Battalion of the 38th Infantry Regiment.

Fig. 6 depicts an integrated view of information related to a casualty in the person perspective of the WarSampo portal. On the left side is a search interface for finding people by name, and on the right is information about the currently selected person. Basic information about a person (e.g., name, birth and death dates and places, occupation, marital status, military rank with promotion dates if available) is displayed on the top. After that, thumbnails of the linked photographs involving the person are shown. By clicking the thumbnails the user can explore the higher resolution versions of the photographs and their captions. Below the photographs are war time events of the person, his military units, military ranks, municipalities where he is known to have been and a link to a Wikipedia page about this person.

[21] http://www.sotasampo.fi/en/units/?uri=http://ldf.fi/warsa/actors/actor_972

For some people, there is also a biography text from the National Biography of Finland shown on the page, and possibly further information like linked magazine articles.

In order to get further context for the person examined, the user can browse the army units the person belonged to during the war, and places related to his life events (e.g., birth and death municipalities on historical and contemporary maps). This way the user is able to track the person's participation in the war by investigating the movements of his army units and the durations of the battles the units fought.



Persons

Search for known persons from the past Finnish wars by writing their name in the text input below and/or selecting a person from the list below. Information regarding the person and recommended links will appear on the right.

niilo korhonen

Niilo Antero Korhonen
Niilo Ensio Korhonen
Niilo Juhani Korhonen
Niilo Juho Korhonen

Niilo Juhani Korhonen

15.12.1915 Nilsiä - 26.9.1943

Kuolinsyy muu kuin sodasta välittömästi aiheutunut (sairaus, tapaturma). Buried in Kuopio

Profession: opiskelija

Marital status: naimisissa, 0 children

Served in 9./JR 10

Luutnantti

Source: Mannerheim-ristin ritarit. http://www.mannerheim-ristinritarit.fi/ritarit

Show information page

7 photos

Units (2)

Jalkaväkirykmentti 10, III Pataljoona, 9. Komppania
Jalkaväkirykmentti 10; Jalkaväkirykmentti 10

Ranks (1)

Luutnantti

Places (3)

Nilsiä
Kuopio
Kuopio

Fig. 6. A screenshot of the person perspective of the WarSampo portal depicting linked information related to a casualty.

6 Discussion

War history data is usually scattered in many isolated silos and may exist in totally different formats, e.g., books, paper archives, and databases. In this paper we have examined the benefits and challenges of linking casualty records of war to multiple related datasets, and publishing them as linked data for DH research and applications to use. Two use cases were studied related to supporting DH research and services for the public.

In the future, linking with other WarSampo data will be developed further as new datasets are added to the system. We plan to develop tools for statistical analysis of the data, and collaborate with humanities researchers in studying how linked data and our tooling can help to solve their research problems. A hierarchical occupation ontology is planned to be used and linked to the death records to provide insight into the social status of each casualty. Whether we could take advantage of existing occupation taxonomies, such as the Historical International Standard Classification of Occupations (HISCO) [10], will be explored.

It would be beneficial for research purposes to develop the casualties perspective to allow exporting the data based on facet selections, for use with other applications and visualization tools.

Acknowledgements

Tomi Ahoranta and Jérémie Dutruit from the National Archives contributed in publishing the original casualty database as LOD.

Our research is partially funded through the Finnish Open Science and Research Initiative by the Finnish Ministry of Education and Culture.

References

1. de Boer, V., van Doornik, J., Buitinck, L., Marx, M., Veken, T.: Linking the kingdom: enriched access to a historiographical text. In: Proc. of the 7th International Conference on Knowledge Capture (KCAP 2013). pp. 17–24. ACM (2013)
2. Collins, T., Mulholland, P., Zdrahal, Z.: Semantic Browsing of Digital Collections. In: Proc. of the 4th International Semantic Web Conference (ISWC 2005). pp. 127–141. Springer–Verlag (2005)
3. Crymble, A., Gibbs, F., Hegel, A., McDaniel, C., Milligan, I., Posner, M., Turkel, W.J. (eds.): The Programming Historian. 2 edn. (2015), http://programminghistorian.org
4. Doerr, M.: The CIDOC CRM—an Ontological Approach to Semantic Interoperability of Metadata. AI Magazine 24(3), 75–92 (2003)
5. Graham, S., Milligan, I., Weingart, S.: Exploring big historical data. The historian's macroscope. Imperial College Press (2015)
6. Hyvönen, E., Ikkala, E., Tuominen, J.: Linked data brokering service for historical places and maps. In: Adamou, A., Daga, E., Isaksen, L. (eds.) Proc. of the 1st Workshop on Humanities in the Semantic Web (WHiSe). pp. 39–52. No. 1608 in CEUR Workshop Proceedings, Aachen (2016), http://ceur-ws.org/Vol-1608/#paper-06
7. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets. In: The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers. pp. 226–230. Springer–Verlag (2014)
8. Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J., Mäkelä, E.: WarSampo data service and semantic portal for publishing linked open data about the second world war history. In: The Semantic Web – Latest Advances and New Domains (ESWC 2016). pp. 758–773. Springer-Verlag (2016)

9. Koho, M., Heino, E., Hyvönen, E.: SPARQL Faceter—Client-side Faceted Search Based on SPARQL. In: Troncy, R., Verborgh, R., Nixon, L., Kurz, T., Schlegel, K., Vander Sande, M. (eds.) Joint Proc. of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop. No. 1615, CEUR Workshop Proceedings (2016), `http://ceur-ws.org/Vol-1615/semdevPaper5.pdf`
10. van Leeuwen, M.H.D., Maas, I., Miles, A.: HISCO: Historical international standard classification of occupations. Leuven University Press (2002)
11. Mäkelä, E.: Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text. In: The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers, pp. 424–428. Springer–Verlag (2014)
12. Mäkelä, E., Törnroos, J., Lindquist, T., Hyvönen, E.: World War I as Linked Open Data (2015), `http://seco.cs.aalto.fi/publications/`, submitted for review
13. Tunkelang, D.: Faceted search. Synthesis lectures on information concepts, retrieval, and services, Morgan & Claypool Publishers (2009)

Publication IX

Mikko Koho, Esko Ikkala, Erkki Heino, and Eero 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, October 29–November 3, 2018, Proceedings, Part I*, Marinos Ioannides, Eleanor Fink, Rafaella Brumana, Petros Patias, Anastasios Doulamis, João Martins, and Manolis Wallace (editors), Information Systems and Applications, incl. Internet/Web, and HCI, volume 11196, pages 138–149, ISBN 9783030017613, Springer, Cham, October–November 2018.

Maintaining a Linked Data Cloud and Data Service for Second World War History

Mikko Koho[1][0000−0002−7373−9338], Esko Ikkala[1,2][0000−0002−9571−7260], Erkki Heino[1], and Eero Hyvönen[1,2][0000−0003−1695−5840]

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

Abstract. One of the great promises of Linked Data is to provide a shared data infrastructure into which new data can be imported and aligned with, forming a sustainable, ever growing Linked Data Cloud (LDC). This paper studies and evaluates this idea in the context of the WarSampo LDC that provides a data infrastructure for Second World War related ontologies and data in Finland, including several mutually linked graphs, totaling ca 12 million triples. Two data integration case studies are presented, in which the original WarSampo LDC and the related semantic portal were first extended by a dataset of hundreds of war cemeteries and thousands of photographs of them, and then by another dataset of over 4450 Finnish prisoners of war. As a conclusion, lessons learned are explicated, based on hands-on experience in maintaining the WarSampo LDC in a production environment.

1 Introduction

This paper studies the fundamental process of building the Web of Data [6] by incrementally aggregating and aligning new datasets into a Linked Data Cloud (LDC). The focus is in particular on publishing and using Cultural Heritage Linked Data on the Semantic Web [8].

We first overview previous research related to the problem of maintaining ontologies and linked data. Based on this, a typology of change propagation in interlinked Resource Description Network (RDF)[3] graphs is presented. Two practical case studies are discussed where a new dataset is integrated into the WarSampo LDC [9], which contains a dynamic ontology infrastructure and a collection of Linked Open Data about Finland in the Second World War (WW2). In both cases, change propagation scenarios are discussed, with lessons learned explicated. As a conclusion, guidelines for integrating a new dataset into an LDC are outlined.

The main contribution of this paper is to address the linked dataset maintenance problem on an LDC level. The paper contributes also by explaining how the new datasets can be shown to the end user as new application perspectives

[3] `https://www.w3.org/RDF/`

and through enriching other existing application perspectives with additional data.

WW2 data is of great interest not only to historians, but to potentially hundreds of millions of citizens globally whose relatives participated in the war, creating a global shared trauma. However, data about the WW2 is scattered in various organizations and countries, written in multiple languages, and represented in heterogeneous formats. WarSampo [9] provides a novel infrastructure for publishing WW2 data as LOD. The infrastructure supports integrating new datasets into WarSampo, by extending both the DOs and the MDSs. Published in 2015, WarSampo is to our best knowledge the first large scale system for serving and publishing WW2 LOD on the Web. WarSampo is a part of the global LOD cloud[4], and was awarded with the LODLAM Challenge Open Data Prize in 2017.

The data is served on an open data service[5], which enables anyone to build applications that use the data via standard APIs. The WarSampo semantic portal uses the data service to provide different perspectives to the WW2 LOD as customized web applications. New perspectives can be added in a flexible way to provide views to new data, or to answer new research questions with existing data.

The War Cemetery perspective is an in-use application on the Semantic Web: it was published in November 2017 and got 57 000 users in one week after that. The Prisoners of War perspective will be published later in 2018. In total, the WarSampo data service was used by 130 000 different users through the WarSampo semantic portal[6] in 2017.

2 Related Work

The problem of maintaining ontologies and linked data have been studied extensively, but mostly from a point of view of editing and managing evolving ontologies and data, not on an LDC level as in this paper. Early works on this line of research include, e.g., [11,15]. In [20,3], the problem of managing a set of interlinked hierarchical RDFS thesauri is discussed. Ontology evolution, and the propagation of changes caused by it, has been discussed in [23] and [25].

Umbrich et al. [24] have surveyed solutions to detect, propagate and describe changes in Linked Open Data resources and datasets. Requirements and approaches are studied for different use cases, e.g. link maintenance and vocabulary evolution. These linked data dynamics are explored also in [2,16]. Handling broken links in Linked Data is discussed in [22].

In addition to the global LOD cloud, other LDCs, like the Lexvo [17] and the MIDI LDC [18] have been previously studied.

A framework for integrating heterogeneous OpenCourseWare data repositories into a Linked Data publication is presented in [21]. A framework and tool for

[4] `http://linkeddata.org`

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

[6] `https://sotasampo.fi/en/`

data fusion, conflict resolution, and quality assessment of Linked Data graphs is presented in [19]. Knoblock et al. [12] presented lessons learned in integrating heterogeneous data from 14 museums into a Linked Data publication, harmonizing data with CIDOC CRM[7].

An overview of the WarSampo data service and semantic portal has been presented in [9]. A core dataset of WarSampo, the casualties of war, and its application in digital humanities research is presented in [14]. Using the war cemetery data in prosopographical research is discussed in [10]. Overview of the Prisoners of War case study with preliminary results have been published [13], with a comparison of different online publishing approaches. Named entity linking in WarSampo was studied in [7]. This paper provides a new view to this line of research from an LDC management point of view.

3 Anatomy and Maintenance of Linked Data Clouds

An LDC consists of a set of graphs. Data is interlinked across graphs by mappings and direct references to URIs in other graphs. We differentiate the graphs into two major categories based on their usage: *metadatasets* (MDS) and *domain ontologies* (DO). MDSs describe objects or other things in an application domain in terms of a metadata schema [4], such as Dublin Core or CIDOC CRM. Collection metadata in libraries, museums, and archives, or their harmonized aggregated versions are typical examples of MDSs. DOs define the basic concepts used in populating the MDSs and are shared by them. DOs include, e.g., ontologies for subject matter concepts (keyword thesauri), places, people, times, and events. The generic, domain independent structure and semantics of DOs and MDSs are defined by a set of shared domain independent vocabularies, such as RDF(S), SKOS, and OWL. Data linking in an LDC is based on making references to shared domain independent vocabularies, domain specific DOs, and mappings.

We call a set of DOs used for populating a set of MDSs in an application domain the *ontology infrastructure*. In many cases, DOs, MDSs, mappings, and domain independent vocabularies are published as one homogeneous triple mass. If there is no separation of DOs and MDSs into graphs, the distinction between them can be vague. An example of this is DBpedia[8], in which resources are separated by namespaces, but this distinction is insufficient, since typically one graph can use a variety of different namespaces. A key observation underlying this paper is that from a data management point of view, DOs, MDSs, and mappings are different from each other, and it makes sense to keep them separate in order to support different kind of maintenance operations.

An important property of a graph is *independence*: we define a graph independent if it does not make a reference to (i.e. links to) resources in other graphs. For example, SKOS keyword thesauri are often independent DOs mak-

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

[8] `http://dbpedia.org`

ing only `skos:broader/ narrower/related` references to concepts within the same concept scheme.

A Typology of Change Propagation A graph can change through changes in its resources. The following three change types are the most fundamental: 1) *Addition.* A new resource is added into the graph. 2) *Modification.* A resource is modified in terms of its properties. 3) *Removal.* A resource is removed from the graph. Based on the primitive changes, more complex changes can be modeled as sequences of more primitive ones, such as moving a resource from a graph into another. The primitive changes may occur in a DO or an MDS, and may have an effect in related DOs or MDSs [23]. We have identified the following four principal cases of change propagation needs between graph types. Here the notation $X \rightarrow Y$ means that a change in a graph X creates a potential need for a change in a graph Y that makes a reference to X.

1. **DO→MDS**. In all cases, linkage based on probabilistic entity linking, from an MDS to the DO, needs to be revalidated. **Addition:** An addition in the DO usually doesn't create a need for change propagation to MDSs. However, when a new DO resource is introduced in an MDS, the linkage from the MDS to the DO is broken since the new resource is not there in the DO. **Modification:** no additional effect. **Removal:** The MDSs can get corrupted by having URI references to removed URIs. The affected MDSs need to be fixed.
2. **DO→DO**. If the changed DO is independent, there are no change propagation needs. Otherwise change propagation is needed as in case DO→MDS.
3. **MDS→DO**. **Addition:** If DOs cover the values used by the MDS, there is no effect. Otherwise the DOs may need to be updated accordingly. **Modification:** usually no effect. If a new value not in a DO would be needed as a property value in the MDS, the DO may need to be updated accordingly. **Removal:** no effect, unless a DO makes a reference to the MDS. This may happen, e.g., when an event ontology makes a reference to an artifact collection database.
4. **MDS→MDS**. Changes between MDSs are propagated as in MDS→DO.

Practical examples of the change propagation scenarios are presented in the use cases in Sections 5 and 6.

4 Maintaining the WarSampo Linked Data Cloud

Creating the WarSampo ontology infrastructure has been a dynamic process, involving several people working with up to seven datasets at the same time. The metadatasets and domain ontologies have been constantly evolving, which often causes existing entity matching to be invalidated.

Fig. 1 shows the main MDSs and DOs of WarSampo, after the data model changes caused by the case studies presented in this paper. Each MDS and DO shows the number of individual entities belonging to the corresponding class(es). The arrows depict the direction of linkage, which is normally from the MDSs

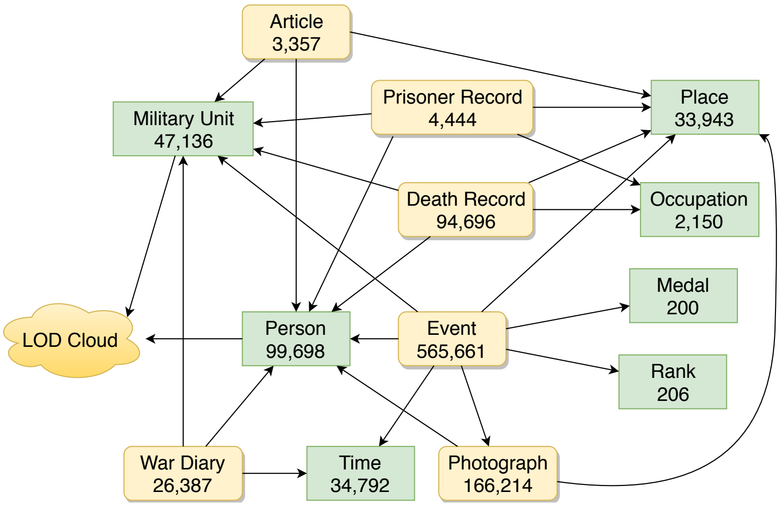


Fig. 1. The main metadatasets (yellow, rounded corners) and domain ontologies (green rectangular boxes) of WarSampo.

towards the DOs that have been used in annotating the entities. There is also linkage to the global LOD cloud.

The WarSampo LDC is centrally maintained, even if it is based on data from distributed sources. This means that DOs can adapt to changes that are needed when integrating new datasets into the LDC, and should do so to better represent their domain. The ontology infrastructure is extended as needed.

Maintaining the WarSampo LDC is different from maintaining the global LOD cloud, where `owl:sameAs` and related mappings between datasets are created, but changes are seldom propagated across the datasets. This is not feasible in our case, since a new piece of information in one graph of the service may require changes in other graphs, too. For example, if a new place or a person is introduced in a new or existing MDS, the Place DO or the Person DO has to be extended before the new data can be aligned.

Integrating data into a dynamic LD environment is challenging. As a DO becomes more complete, covering its domain more accurately, MDSs using that DO may need to redo their entity linking process, to get more accurate linkage. Failing to do so can cause structurally or semantically erroneous annotations [22] to be used. WarSampo employs plenty of probabilistic entity linking, e.g., to the Person, Place and Military Unit DOs, in which the DO is not expected to cover all of the information about its domain. The usual case of change propagation in the WarSampo context relates to the invalidation of the entity linking.

Because of the complex change propagations in the dynamic LDC, maintaining the DOs and MDSs directly in RDF format is too laborious and error-prone to implement in practice, especially in a way that a domain expert with little Linked Data expertise could make changes. This is especially true in the case of person instances as they are linked, directly or indirectly, to everything in WarSampo. Modeling even just the basic information of a person entails, e.g., multiple events, such as birth, death, promotions, an so on. It was decided that the domain experts directly maintain the datasets in their native formats (usually spreadsheet files), which can then be easily transformed and integrated into WarSampo, as needed.

5 Case 1: War Cemeteries

In our first case study, a war cemetery dataset was produced and integrated into the WarSampo LDC [10]. Since Finnish soldiers who perished in WW2 were transported back to their hometown for burial whenever it was possible, the local cemetery is a natural starting point for studying the common characteristics and events of the residents of one's hometown in the turmoil of the war.

Starting Point & Source Data A complete listing of war cemeteries in Finland was not available, but the Casualties MDS, that was previously integrated into WarSampo, includes the name of the cemetery and/or the municipality in which the person is buried. However, the lack of uniform naming conventions and missing coordinates of the cemeteries made it difficult to locate them and to specify the people buried there.

In 2016–2017 the Memorial Foundation of the Fallen and the Central Organization of Finnish Camera Clubs (Suomen Kameraseurojen Liitto ry, SKsL) carried out a project called "War Cemeteries in Finland". Its goal was to locate, photograph, and collect data about all known war cemeteries in Finland. In total 615 war cemeteries were found, accompanied by 2500 photographs.

Workflow A representative of the SKsL manually harmonized the data entry sheets and filenames of the photographs sent by the camera clubs, and organized them into one table. Finally the table was converted into WarSampo compatible RDF by using a Python data processing pipeline[9], which 1) handles the matching of existing cemetery names found in the WarSampo death records to cemetery names in the source table, 2) creates new URIs for cemeteries not found in WarSampo, and 3) creates photograph and photography instances according to the WarSampo data model. Whenever there is a need to update the cemetery data, the source table can be edited and the data processing pipeline can be run again to produce new RDF files for WarSampo.

To avoid errors in the data integration, the "War Cemeteries in Finland" project was instructed to start with the same cemetery name listing that was used when the death records were collected into a database. A challenge here was that some of the cemeteries mentioned in the Casualties MDS were unambiguous.

The structure of the project's output table was agreed on beforehand, so that information about one individual cemetery was gathered in one row, with values separated on columns, easing the RDF conversion process. The cemetery data processing pipeline was run multiple times in order to enhance data quality, and

[9] https://github.com/SemanticComputing/cemeteries-csv2rdf

a listing of spelling errors, missing photograph files, etc. was sent back to SKsL for making manual corrections to their table.

The cemetery data was integrated into the Place DO, and the cemetery photograph data into the Photograph MDS. The photographs are generally linked to the photography places via a photography event, which has created the photograph. However, photographs of war cemeteries represent the cemeteries, which are modeled as a subclass of the place class. Some military units and people are mapped to entities in the global LOD Cloud, i.e., Wikidata and DBpedia.

Change Propagation With regard to maintenance, the basic data about the cemeteries remains independent in the Place DO. If the cemeteries in Place DO change, the linkage from the death records in the Casualties MDS need to adjust for the change as according to the DO→MDS case in Section 3.

However, the information about the people buried in a cemetery is stored in the Casualties MDS which makes references to the Place DO. Thus, the changes in Casualties or Photograph MDS related to the cemeteries must be propagated to the Place DO according to the MDS→DO case in Section 3.

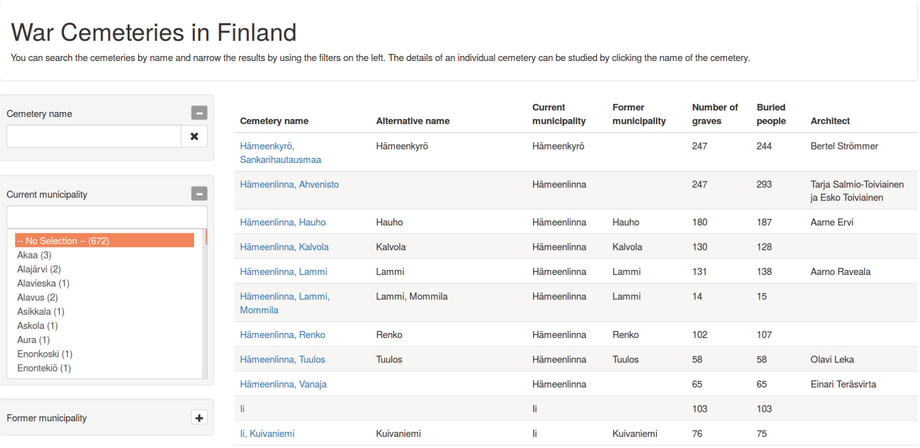
Semantic Portal Changes The new War Cemeteries Perspective[10] showcases how the integration of cemetery data enriches the existing WarSampo data and vice versa. The perspective has been developed to gain new insights from the casualties data based on the community-level aspect provided by the cemeteries. This approach is useful, because there is not enough data about the casualties to construct detailed life stories of individual soldiers as biographies, but the amount of individuals is large enough to study the data as groups of people using, for example, visualizations.

The user interface of the Cemetery perspective is presented in Fig. 2. The user can browse all cemeteries, or search the cemeteries by name and narrow the results by using the filters on the left. The results can be viewed as a table with basic information about the cemeteries, or on a map which provides a global view of the cemeteries.

A concrete example of the data integration results can be seen in Fig. 2, where the ”Number of graves” column is based on the data of the “War Cemeteries in Finland” project, whereas the ”Buried people” column shows the total number of death records (collected in the 1980s) that make a reference to the cemetery. The numbers are equal only with 27 % of the cemeteries although ideally they should be equal with every cemetery. This gives valuable insight to the data providers to set the records straight.

When the user clicks the name of a cemetery, an information page opens, showing basic information, photographs, and various visualizations based on the property values of the buried people.

[10] `https://www.sotasampo.fi/en/cemeteries/`



War Cemeteries in Finland

You can search the cemeteries by name and narrow the results by using the filters on the left. The details of an individual cemetery can be studied by clicking the name of the cemetery.

Cemetery name

Current municipality

-- No Selection -- (672)
Akaa (3)
Alajärvi (2)
Alavieska (1)
Alavus (2)
Asikkala (1)
Askola (1)
Aura (1)
Enonkoski (1)
Enontekiö (1)

Former municipality

Cemetery name	Alternative name	Current municipality	Former municipality	Number of graves	Buried people	Architect
Hämeenkyrö, Sankarihautausmaa	Hämeenkyrö	Hämeenkyrö		247	244	Bertel Strömmer
Hämeenlinna, Ahvenisto		Hämeenlinna		247	293	Tarja Salmio-Toiviainen ja Esko Toiviainen
Hämeenlinna, Hauho	Hauho	Hämeenlinna	Hauho	180	187	Aarne Ervi
Hämeenlinna, Kalvola	Kalvola	Hämeenlinna	Kalvola	130	128	
Hämeenlinna, Lammi	Lammi	Hämeenlinna	Lammi	131	138	Aarno Raveala
Hämeenlinna, Lammi, Mommila	Lammi, Mommila	Hämeenlinna	Lammi	14	15	
Hämeenlinna, Renko	Renko	Hämeenlinna	Renko	102	107	
Hämeenlinna, Tuulos	Tuulos	Hämeenlinna	Tuulos	58	58	Olavi Leka
Hämeenlinna, Vanaja		Hämeenlinna		65	65	Einari Teräsvirta
Ii		Ii		103	103	
Ii, Kuivaniemi	Kuivaniemi	Ii	Kuivaniemi	76	75	

Fig. 2. Cemetery search in the WarSampo cemetery perspective.

6 Case 2: Prisoners of War

Some 4450 Finnish soldiers were captured as prisoners of war (POW) in WW2 by the Soviet Union. This case study concerns integrating the POW data into the WarSampo LDC.

Starting Point & Source Data The POW dataset was originally published in a book [1]. Recently, the dataset has been extensively extended, cleaned, and validated by domain experts. A collaboration was set up to publish the data as part of the WarSampo, which was chosen as the primary data publication platform by the stakeholders, which include the National Archives of Finland, and the Association for Cherishing the Memory of the Dead of the War.

The core of the dataset is a register of the Finnish prisoners of war in WW2. The register is formatted as a spreadsheet file, with additional spreadsheet files presenting data about POW camps and hospitals, as well as the primary data sources. The POW dataset contains sensitive information about the individual soldiers, some of whom are still alive. There is an ongoing process to evaluate what information can be published, by the legal experts at the National Archives of Finland. The data will be published in the autumn 2018, at which point the privacy issues should be resolved.

Workflow The data formatting evolved as a collaboration between the domain experts maintaining the original dataset, and the WarSampo team of Linked Data experts. A data processing pipeline was created[11], that handles data transformation, validation, linking, and harmonization. The pipeline transforms the spreadsheets into RDF, mapping the spreadsheet columns to RDF properties, with possibly multiple values per property, and containing annotations for primary information sources. Automatic linking processes then link the

[11] `https://github.com/SemanticComputing/WarPrisoners`

records to WarSampo DOs of military ranks, units, occupations, people, and places.

The prisoner records were modeled in a way similar to the previously published Casualties MDS [14], and they share common super classes and properties. However, the process workflow was different: the casualty data was received as a static data dump, whereas the POW dataset was constantly evolving during the project.

The original POW register is maintained in spreadsheet format, which can be easily integrated into WarSampo with our automated transformation process when the spreadsheet is updated, provided that the structure stays the same.

For most of the original data, the spreadsheet format is a natural way to represent the information, with each row of the POW register expressing information about one individual soldier, and each column representing a different property of a soldier, like his name, occupation, and date of capture.

As the data comes from multiple sources that can have contradictory information, there is a need to collect all different values for a single property, along with references to the primary data sources. For this purpose, a special cell data format is used that enables to present multiple values and source references in the spreadsheet. The cell formatting is validated during the data transformation process. Also other simple data validation rules are applied to find anomalies during data conversions.

Change Propagation The POW data introduces the main MDS of POWs, and a DO of war-time occupations. The WarSampo person DO is updated with about 3,000 new person instances. POW camps and hospitals are modeled as part of the Place DO.

The original dataset contains source references for separate pieces of information, which are used in the RDF data model by employing RDF reification for the prisoner records. This is a standard approach to modeling this kind of provenance information on an RDF triple level.

The DOs of military ranks, military units, places (e.g. municipalities, camps and hospitals), occupations and persons provide values for populating the POW MDS. Their linking uses probabilistic entity linking, while also original values are stored as literals. All changes in the DOs would require repeating the corresponding entity linking process as according to the case DO→MDS in Section 3. I.e. if a new understanding about the historical war-time Occupation DO (cf. Fig. 1) cause two occupations to be merged into one, resulting in the removal of the obsolete one, any linking to the obsolete resources need to be updated.

Adding a new property value in the MDS can propagate the change to related DOs, if the value doesn't exist there (cf. Section 3, case MDS→DO). For example, the new value could be a new military rank or a new occupation. When a new POW record is added to the registry, the changes will propagate to the Person DO, either through the linking to an existing person, in which case the person instance is enriched, or through the creation of a new person instance.

The POW records are mapped to the Person DO using probabilistic record linkage [5], where each POW's information is compared with the information in the WarSampo person instances to find matches that have high enough similarity. As the record linkage needs to be able to adapt to changing input dataset, as well as to the changes in the Person DO, a machine learning approach was used, which employs logistic regression based on weighted comparisons of a set of predefined attributes. The weights are calculated based on training data, which is initially acquired from a previous, simpler record linkage implementation, based on manually defined fuzzy matching, and updated manually during linkage iterations. With the machine learning approach, the entity linking process automatically adapts to changes in the POW MDS and Person DO. The linking process needs to be redone when the POW MDS changes.

New person instances are then created for the unlinked POW records and added into the Person DO. With the probabilistic linkage, it is possible that a record is not mapped because there is not enough information about either the POW record, or the person instance, to create a mapping between them. Modifying the information in either the MDS or the DO means that the whole record linkage process should be redone.

Semantic Portal Changes A new application perspective has been added to WarSampo to explore, analyze, and visualize the information contained in the POW metadataset. The perspective is similar to the earlier casualties perspective, which is used to show information from the death records to the user.

In addition, integrating the prisoners of war data into WarSampo has caused several necessary changes to other parts of the semantic portal. Allowing multiple values for properties with provenance data changes how the information can be presented in a person's home page and how to visualize the data. People's home pages in WarSampo were updated to show information combined from multiple sources (death records, prisoner records, Wikipedia) with source information next to each piece of information.

7 Conclusions

A key lesson learned in our work is that one should make all data transformations and linking into **repeatable, automated processes** to be able to handle change propagation automatically. In the early stages of building WarSampo, the importance of this was not obvious, and for some early WarSampo datasets, the transformation processes were never completely automated. Automating them now would require considerable effort because the datasets have gone through undocumented processes that are not easily repeatable.

The transformation processes should be built using a modular structure, to make the processes **maintainable**, and to enable the reuse of code for other data integration. In a dynamic LDC, the entity linking processes need to be able to **adapt** to common changes in all of the graphs.

Maintenance of an LDC using a complex data model, such as CIDOC CRM, is difficult natively in RDF format. For complex DOs and MDSs, it is easier to update the data in simpler formats, such as Dublin Core, and maintain the transformation processes that build the graphs of the LDC. The complexity of the transformation processes grows as they need to handle the creation or updating of missing or uncertain resources in incomplete DOs shared by multiple MDSs. Simple, independent DOs (e.g. military units, occupations) can be maintained directly in RDF format, whereas more complex DOs like Persons require a different approach.

DOs differ from each other by nature. For example, covering and disambiguating all military ranks is clearly a simpler task than performing the same task with all wartime places. In general, it is not realistic to assume that the DOs completely cover their domain.

Integrating data into a LDC is more laborious than simpler ways of publishing the data in independent data silos. However, the result is an interlinked knowledge base, a Linked Data Cloud, where the interlinked graphs enrich each other, creating a whole that is greater than the sum of its parts.

Acknowledgements Our work was funded by the Association for Cherishing the Memory of the Dead of the War, the Memory Foundation for the Fallen, the Finnish Ministry of Education and Culture, and the Academy of Finland. The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

1. Alava, T., Frolov, D., Nikkilä, R.: Rukiver. Suomalaiset sotavangit Neuvostoliitossa. Helsinki: Edita (2003)
2. Auer, S., Dalamagas, T., Parkinson, H., Bancilhon, et al.: Diachronic linked data: towards long-term preservation of structured interrelated information. In: Proceedings of the First International Workshop on Open Data. pp. 31–39. ACM (2012)
3. Frosterus, M., Tuominen, J., Pessala, S., Hyvönen, E.: Linked Open Ontology cloud: managing a system of interlinked cross-domain light-weight ontologies. International Journal of Metadata, Semantics and Ontologies **10**(3), 189–201 (2015)
4. Gartner, R.: Metadata. Shaping Knowledge from Antiquity to the Semantic Web. Springer–Verlag (2016)
5. Gu, L., Baxter, R., Vickers, D., Rainsford, C.: Record linkage: Current practice and future directions. CSIRO Mathematical and Information Sciences Technical Report **3**, 83 (2003)
6. Heath, T., Bizer, C.: Linked Data: Evolving the web into a global data space. Synthesis Lectures on The Semantic Web: Theory and Technology, Morgan & Claypool Publishers, Palo Alto, USA (2011)
7. 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: Proceedings, Language, Technology and Knowledge (LDK 2017). pp. 120–133. Springer-Verlag (June 2017)
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, USA (2012)
9. Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J., Mäkelä, E.: WarSampo data service and semantic portal for publishing linked open data about the second world war history. In: The Semantic Web – Latest Advances and New Domains (ESWC 2016). pp. 758–773. Springer-Verlag (2016)

10. Ikkala, E., Koho, M., Heino, E., Leskinen, P., Hyvönen, E., Ahoranta, T.: 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 (October 2017)
11. Klein, M.: Change Management for Distributed Ontologies. Ph.D. thesis, Free University, Amsterdam (2004)
12. Knoblock, C.A., Szekely, P., Fink, E., Degler, D., Newbury, D., Sanderson, R., Blanch, K., Snyder, S., Chheda, N., Jain, N., et al.: Lessons Learned in Building Linked Data for the American Art Collaborative. In: International Semantic Web Conference. pp. 263–279. Springer (2017)
13. Koho, M., Heino, E., Ikkala, E., Hyvönen, E., Nikkilä, R., Moilanen, T., Miettinen, K., Suominen, P.: Integrating Prisoners of War Dataset into the WarSampo Linked Data Infrastructure. In: Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018). CEUR Workshop Proceedings (March 2018), `http://www.ceur-ws.org/Vol-2084`, vol 2084
14. 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 2017 Satellite Events. pp. 369–383. Springer-Verlag (2017), `https://doi.org/10.1007/978-3-319-70407-4_45`
15. Maedche, A., Motik, B., Stojanovic, L., Studer, R., Volz, R.: An infrastructure for searching, reusing and evolving distributed ontologies. In: Proc. of the twelfth international conference on World Wide Web. pp. 439–448. ACM Press (2003)
16. Meimaris, M., Papastefanatos, G., Pateritsas, et al.: Towards a Framework for Managing Evolving Information Resources on the Data Web. In: Proceedings of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data. CEUR Workshop Proceedings (March 2014), vol 1151
17. de Melo, G.: Lexvo. org: Language-related information for the linguistic linked data cloud. Semantic Web **6**(4), 393–400 (2015)
18. Meroño-Peñuela, A., Hoekstra, R., Gangemi, A., Bloem, P., de Valk, R., Stringer, B., Janssen, B., de Boer, V., Allik, A., Schlobach, S., et al.: The MIDI linked data cloud. In: International Semantic Web Conference. pp. 156–164. Springer (2017)
19. Michelfeit, J., Knap, T., Nečaskỳ, M.: Linked data integration with conflicts. arXiv preprint arXiv:1410.7990 (2014)
20. Pessala, S., Seppälä, K., Suominen, O., Frosterus, M., Tuominen, J., Hyvönen, E.: MUTU: An Analysis Tool for Maintaining a System of Hierarchically Linked Ontologies. In: ISWC 2011 - Ontologies come of Age Workshop (OCAS). CEUR Workshop Proceedings (2011), vol 809
21. Piedra, N., Tovar, E., Colomo-Palacios, R., Lopez-Vargas, J., Alexandra Chicaiza, J.: Consuming and producing linked open data: the case of Opencourseware. Program **48**(1), 16–40 (2014)
22. Popitsch, N.P., Haslhofer, B.: Dsnotify: handling broken links in the web of data. In: Proceedings of the 19th international conference on World wide web. pp. 761–770. ACM (2010)
23. Stojanovic, L., Maedche, A., Motik, B., Stojanovic, N.: User-driven ontology evolution management. In: International Conference on Knowledge Engineering and Knowledge Management. pp. 285–300. Springer (2002)
24. Umbrich, J., Villazón-Terrazas, B., Hausenblas, M.: Dataset dynamics compendium: A comparative study (2010)
25. Zablith, F., Antoniou, G., d'Aquin, M., Flouris, G., Kondylakis, H., Motta, E., Plexousakis, D., Sabou, M.: Ontology evolution: a process-centric survey. The knowledge engineering review **30**(1), 45–75 (2015)

This thesis explores the use of Semantic Web technologies for representing and modeling heterogeneous military historical information as Linked Data. Harmonization and integration of military historical data from distributed sources are studied, while also investigating how to search, browse, analyze, and visualize the resulting Linked Data on web-based user interfaces. Maintenance of the highly interlinked set of graphs exposes new challenges and a solution to tackle them is presented. These topics are studied in the context of building the WarSampo information system, which contains a knowledge graph of ca. 14 million triples and the popular web-based WarSampo portal for accessing the information contained in the knowledge graph.

Aalto-DD 71/2020

9 789526 038681

ISBN 978-952-60-3868-1 (printed)
ISBN 978-952-60-3869-8 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Computer Science
www.aalto.fi