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

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

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/

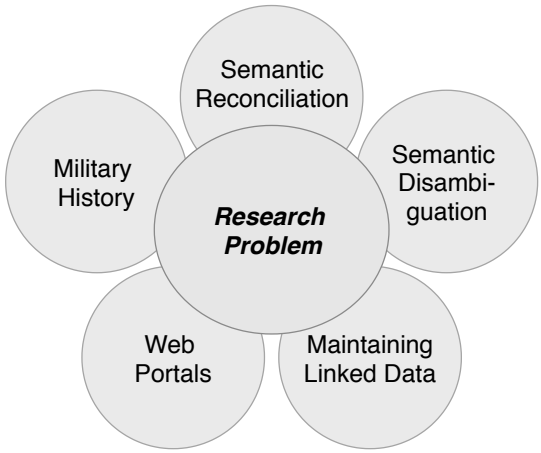

Figure 1.1. A diagram showing the research areas of the thesis, and depicting the combination of the research areas as the research gap of the thesis.

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].

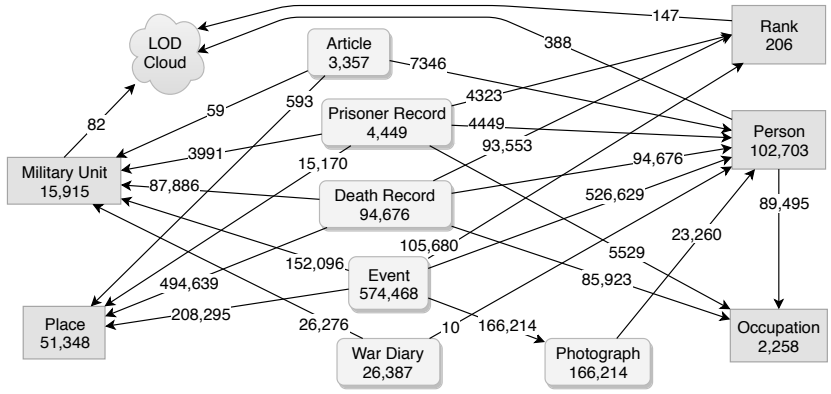

Figure 3.1. The core classes with instance counts and linkage between class instances.

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 mili-

tary 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.

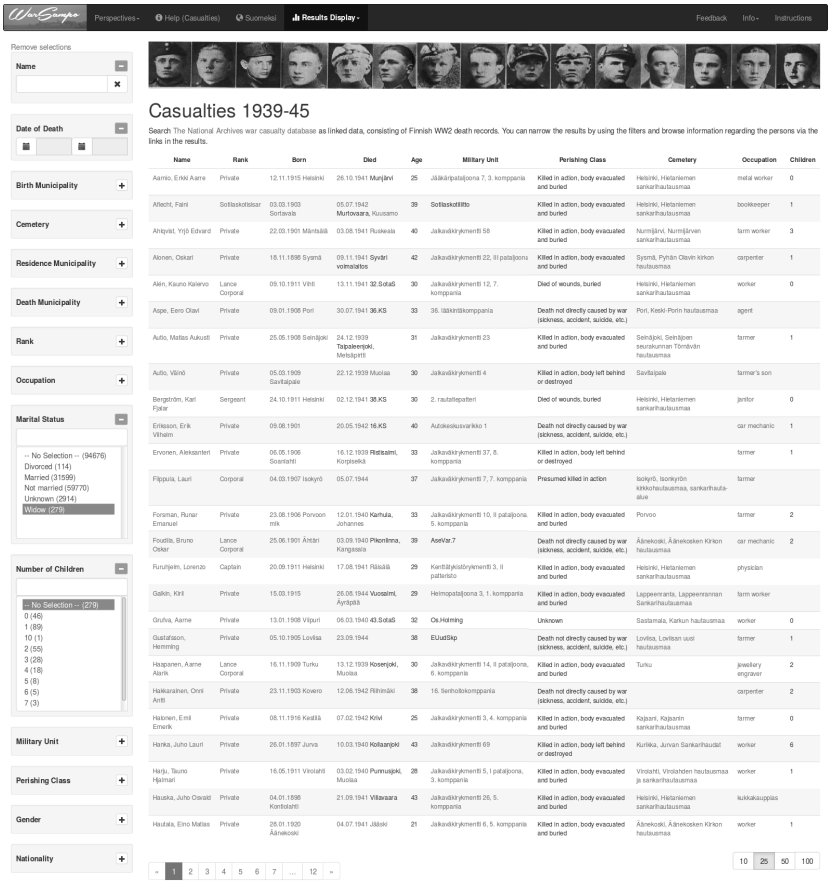

Figure 3.2. The faceted search interface of the casualties perspective with one selected facet. The left side contains the facets, displaying available categories and the amount of death records for each selection category. Death records matching the current facet selections are shown as a table.

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.

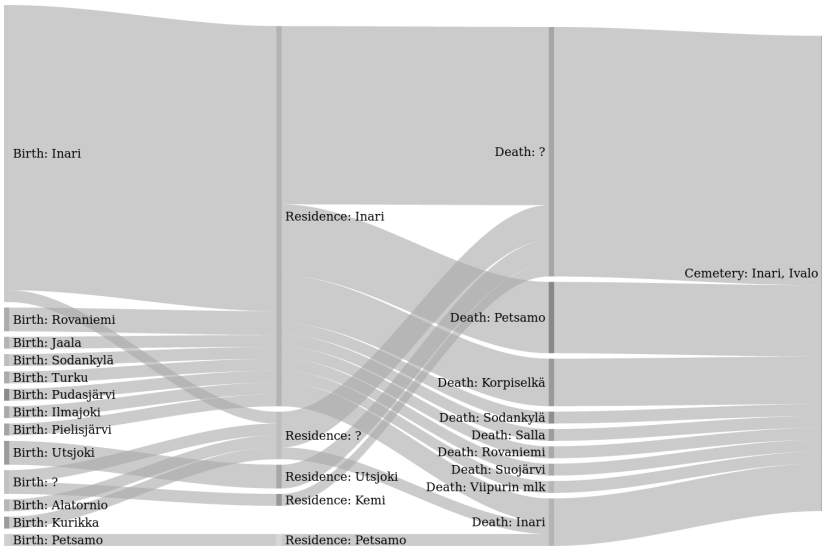

Figure 3.3. Life paths of 40 soldiers buried in the cemetery of Inari in Ivalo.

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. Musteață, 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.

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

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

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS