Computer Science

From Text to Knowledge: Methods, Tools, and Applications for Digital Humanities Based on Linked Data

Minna Tamper

Aalto University

DOCTORAL THESES

Aalto University publication series
DOCTORAL THESES 19/2023

From Text to Knowledge: Methods, Tools, and Applications for Digital Humanities Based on Linked Data

Minna Tamper

A doctoral thesis completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall T2 of the school on 31 March 2023 at 12 noon.

Aalto University
School of Science
Computer Science
Semantic Computing Research Group

Supervising professor
Professor Eero Hyvönen, Aalto University, Finland

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

Preliminary examiners
Professor Veronika Laippala, University of Turku, Finland
Lecturer (PhD) John P. McCrae, University of Galway, Ireland

Opponent
Professor Veronika Laippala, University of Turku, Finland

Aalto University publication series
DOCTORAL THESES 19/2023

ISBN 978-952-64-1150-7 (printed)
ISBN 978-952-64-1151-4 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-1151-4

Images: Cover image by Minna Tamper, 2015, CC BY-NC-ND 4.0

Unigrafia Oy
Helsinki 2023

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

A? Aalto University

Abstract

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

Author
Minna Tamper

Name of the doctoral thesis
From Text to Knowledge: Methods, Tools, and Applications for Digital Humanities Based on Linked Data

Publisher School of Science

Unit Computer Science

Series Aalto University publication series DOCTORAL THESES 19/2023

Field of research Computer science

Manuscript submitted 31 August 2022 **Date of the defence** 31 March 2023

Permission for public defence granted (date) 17 January 2023 **Language** English

□ **Monograph** ☒ **Article thesis** □ **Essay thesis**

Abstract

The digitization of Cultural Heritage collections has enabled the use of computational methods such as Natural Language Processing (NLP) on textual collections. These methods have been used widely in Digital Humanities (DH) to study digitized contents with automated processes. The Semantic Web and linked data technologies have been applied to describe document collections and their metadata in library and museum collections. They provide infrastructure for connecting different collections by linking them using shared vocabularies that describe metadata values and fields.

Linked data is also used in Finnish museum and library collections. It is commonly used to modeling document metadata, such as author, or title of a piece of work. Also, the content of a document in a collection is usually described using manually assigned keywords. Other information about the content is often scarce and finding documents related to an actor can be laborious. This thesis studies and presents novel models, methods, and tools for transforming and enriching document collections automatically to linked data. Linked data technology helps to link together documents of a collection based on their metadata, e.g., author, or publisher. It can be also used to link documents based on information extracted about the content, such as actors mentioned in text.

The aim of this thesis is to study how the NLP methods and linked data can be used to study digitized document collections, such as biographies. Research in this thesis is conducted by designing, implementing, and evaluating proof-of-concept systems, tools, and data for real life use cases. The research follows the principles of the design science and action research.

The thesis presents a toolkit that can be used to model, transform, and enrich biographical text document collections to linked data to improve collection's information retrieval and interoperability internally and with other collections. The data model for describing text document collection's content and features, e.g., keywords and mentioned names, creates a foundation for building intelligent services based on the linked data such as network or linguistic analysis. These services can be used to visualize the interlinked data by showing the relations between themes or actors. In addition, the linked-data-based datasets can be used as an input for NLP tools to create data analytical visualizations and applications. This approach can be also used to evaluate the quality and content of text document collections for DH research. The prototypes created for data transformation, enrichment, and information visualization can be also applied to other document collections.

Keywords natural language processing, linked data, data modelling, biographies

ISBN (printed) 978-952-64-1150-7 **ISBN (pdf)** 978-952-64-1151-4

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

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2023

Pages 218 **urn** http://urn.fi/URN:ISBN:978-952-64-1151-4

Aalto-yliopisto

Tiivistelmä

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

Tekijä
Minna Tamper

Väitöskirjan nimi
From Text to Knowledge: Methods, Tools, and Applications for Digital Humanities Based on Linked Data

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietotekniikan laitos

Sarja Aalto University publication series DOCTORAL THESES 19/2023

Tutkimusala Tietotekniikka

Käsikirjoituksen pvm 31.08.2022 **Väitöspäivä** 31.03.2023

Väittelyluvan myöntämispäivä 17.01.2023 **Kieli** Englanti

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

Tiivistelmä

Kulttuuriperintökokoelmien digitalisointi on avannut tekstiaineistot tietokoneavusteisille menetelmille, kuten luonnollisen kielen käsittelylle. Digitaalisissa ihmistieteissä näitä menetelmiä käytetään laajalti digitoitujen aineistojen ja niiden sisällön tutkimiseen automaattisien prosessien avulla. Semanttisen Webin ja linkitetyn datan teknologiaa hyödynnetään kirjastojen ja museoiden asiakirjakokoelmien sekä niiden metadatan kuvailussa. Ne luovat infrastruktuurin, jonka avulla voidaan yhdistää erilaisia kokoelmia käyttämällä niiden linkityksessä jaettuja sanastoja kuvaamaan aineistojen metadatan arvoja ja kenttiä.

Suomessa linkitetyn datan infrastruktuureja hyödynnetään muun muassa museoissa ja kirjastoissa. Useimmiten sitä käytetään mallintamaan asiakirjojen metadataa, kuten tekijä tai teoksen nimi. Tämän lisäksi tekstiaineistokokoelman teosten sisältöä kuvaillaan usein manuaalisesti tuotetuilla asiasanoilla. Muu informaatio sisällöstä voi olla niukkaa ja teosten löytäminen esimerkiksi sisällössä esiintyvän toimijan perusteella voi olla työlästä. Tässä työssä tutkitaan ja esitellään uusia tietomalleja, työkaluja, ja menetelmiä muuntamaan ja rikastamaan tekstiaineistoja linkitetyksi dataksi. Linkitetyn datan avulla voidaan yhdistää tekstikokoelmien asiakirjat toisiinsa metadatan, kuten tekijän tai kustantajan, perusteella. Sen avulla voidaan myös yhdistää asiakirjoja toisiinsa louhimalla informaatiota sisällöstä, kuten siinä mainitut toimijat.

Tämän työn tavoitteena on tutkia kuinka luonnollisen kielen käsittelyn menetelmien ja linkitetyn datan periaatteiden avulla voidaan tutkia digitoituja tekstidokumenttikokoelmia, kuten biografioita. Tutkimus toteutetaan suunnittelemalla, toteuttamalla, ja arvioimalla prototyyppisovelluksia, työkaluja, ja data-aineistoja todellisen elämän käyttötapauksille. Tämä tutkimus noudattaa suunnittelutieteiden ja toimintatutkimuksen metodologioiden periaatteita.

Tässä työssä esitellään ohjelmistoja, jota voidaan soveltaa biografisten tekstiasiakirjakokoelmien mallinnukseen, muuntamiseen, ja rikastamiseen linkitetyksi dataksi. Näin voidaan parantaa kokoelman teosten keskinäistä linkitystä sekä siihen kohdistuvaa tiedonhakua. Tekstiaineistokokoelman ominaisuuksia ja sisältöä, kuten esimerkiksi asiasanat ja henkilöviittaukset, kuvaava tietomalli luo pohjan linkitettyyn dataan perustuville älykkäille sovelluksille, kuten verkosto- tai kielianalyysille. Näiden sovellusten avulla on mahdollista visualisoida linkitetyn datan muodostama verkosto eri toimijoiden ja teemojen välillä. Tämän lisäksi linkitetyn datan infrastruktuuria voidaan käyttää syötteenä luonnollisen kielen käsittelyn sovelluksille, joita voidaan käyttää luomaan data-analyyttisiä visualisointeja ja sovelluksia. Tätä menetelmää voidaan myös käyttää tekstiaineistokokoelmien laadun ja sisällön arvioimiseen digitaalisten ihmistieteiden tutkimusta varten. Prototyyppisovelluksia, joita on luotu tekstiaineistokokoelmien muuntamista, rikastamista ja tiedon visualisointia varten, voidaan myös soveltaa muihin tekstiaineistokokoelmiin.

Avainsanat luonnollisen kielen käsittely, linkitetty data, datan mallinnus, biografiat

ISBN (painettu) 978-952-64-1150-7 **ISBN (pdf)** 978-952-64-1151-4

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

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2023

Sivumäärä 218 **urn** http://urn.fi/URN:ISBN:978-952-64-1151-4

Preface

Little did I know when I started to study computer science in a vocational school that I would end up writing a doctoral dissertation. In fact, my study coordinator at my grammar school tried to convince me to study to become a baker. Back in the day, I spent my time painting portraits, writing literary fiction, working on computer graphics, and playing video games. Regardless, I ended up studying computer science and I gained enough self-confidence to pursue higher education. That confidence, perseverance, and the friends gained on the way led me eventually to the Semantic Computing Research Group (SeCo) at the Aalto University to do my dissertation.

In 2015, I was given an opportunity to do my Master's thesis and later on doctoral dissertation in SeCo. In SeCo I got a new set of colleagues who helped me to adjust to the academic world. Therefore, I'd like to thank all members (past and present) of SeCo for support, guidance, and fruitful discussion to fuel my dissertation project. I would like to thank my supervisor Eero Hyvönen who enabled the project and kept me on track. I'd also like to thank my advisors for helping me and encouraging me to continue on my chosen path. Thanks to Petri Leskinen, Esko Ikkala, Jouni Tuominen, Eetu Mäkelä, Eero Hyvönen and other members of SeCo for inspirational discussions related to computational linguistics, network analyses, and linked data services. I would also like to thank co-authors and colleagues Kimmo Kettunen, Matti La Mela, Kirsi Keravuori, Risto Valjus, and Kasper Apajalahti for collaborations. Thanks also to Aki Hietanen, Saara Packalén, Tiina Husso, and Oili Salminen of the Ministry of Justice, and Risto Talo, Jari Linhala, and Arttu Oksanen of Edita Publishing Ltd for collaboration. Thanks also to Aleksandra Konovalova from the University of Helsinki and Esko Kirjalainen from the Finnish Digital Agency for insightful discussions and to assistant professor Mikko Kivelä for inspirational discussions. For providing useful tools, insight, and support I'd like to thank members of TurkuNLP research group and FIN-CLARIN research consortium.

My work was funded by the Severi project[1] (funded mainly by Business Finland), Anoppi project[2] (funded by the Finland's Ministry of Justice), the EU project In/Tangible European Heritage – Visual Analysis, Curation and Communication (InTaVia)[3], Open Science and Research Initiative of the Finnish Ministry of Education and Culture, the Finnish Cultural Foundation, the Media Industry Research Foundation of Finland, and the Academy of Finland.

Lastly, I would like to thank CSC IT Center for Science for computational resources.

Helsinki, February 1, 2023,

Minna Tamper

[1] https://seco.cs.aalto.fi/projects/severi/

[2] https://seco.cs.aalto.fi/projects/anoppi/

[3] https://seco.cs.aalto.fi/projects/intavia/

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 Minna Tamper, Petri Leskinen, Esko Ikkala, Arttu Oksanen, Eetu Mäkelä, Erkki Heino, Jouni Tuominen, Mikko Koho and Eero Hyvönen. AATOS – a Configurable Tool for Automatic Annotation. In *Language, Data, and Knowledge – First International Conference, LDK 2017, Proceedings, 19 – 20 June 2017, Galway, Ireland*, Jorge Gracia, Francis Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Sebastian Hellmann (editors), Lecture Notes in Computer Science, Volume 10138, pages 276–289, ISSN 1611-3349, DOI 10.1007/978-3-319-59888-8_24, June 2017.

II Erkki Heino, Minna Tamper, Eetu Mäkelä, Petri Leskinen, Esko Ikkala, Jouni Tuominen, Mikko Koho and Eero Hyvönen. Named Entity Linking in a Complex Domain: Case Second World War History. In *Language, Data, and Knowledge – First International Conference, LDK 2017, Proceedings, 19 – 20 June 2017, Galway, Ireland*, Jorge Gracia, Francis Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Sebastian Hellmann (editors), Lecture Notes in Computer Science, Volume 10138, pages 120–133, ISSN 0302-9743, DOI 10.1007/978-3-319-59888-8_10, June 2017.

III Minna Tamper, Petri Leskinen, Kasper Apajalahti and Eero Hyvönen. Using Biographical Texts as Linked Data for Prosopographical Research and Applications. In *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection, Euromed 2018, 29 Oct – 3 Nov 2018, Nicosia, Cyprus*, Marinos Ioannides, Eleanor Fink, Raffaella Brumana, Petros Patias, Anastasios Doulamis, João Martins, Manolis Wallace (editors), Lecture Notes in Computer Science, Volume 11196, pages 125–137, ISSN 0302-9743, DOI 10.1007/978-3-030-01762-0_11, Nicosia, Cyprus, November 2018.

IV Matti La Mela, Minna Tamper and Kimmo Kettunen. Finding Nineteenth-century Berry Spots: Recognizing and Linking Place Names in a Historical Newspaper Berry-picking Corpus. In *Digital Humanities in the Nordic Countries: Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, 6 – 8 Mar 2019, Copenhagen, Denmark*, Costanza Navarretta, Manex Agirrezabal, Bente Maegaard (editors), CEUR Workshop Proceedings, Volume 2364, pages 308–319, ISSN 1613-0073, online `http://ceur-ws.org/Vol-2364/27_paper.pdf`, Aachen, March 2019.

V Minna Tamper, Petri Leskinen and Eero Hyvönen. Visualizing and Analyzing Networks of Named Entities in Biographical Dictionaries for Digital Humanities Research. Accepted for publication in *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019), La Rochelle, France, June 3, 2019*, CICLing, April 2019.

VI Minna Tamper, Petri Leskinen, Jouni Tuominen and Eero Hyvönen. Modeling and Publishing Finnish Person Names as a Linked Open Data Ontology. In *Proceedings of the Third Workshop on Humanities in the Semantic Web (WHiSe 2020) co-located with 15th Extended Semantic Web Conference (ESWC 2020) Heraklion, Greece, June 2, 2020 (online)*, Alessandro Adamou, Enrico Daga, Albert Meroño-Peñuela (editors), CEUR Workshop Proceedings, Volume 2695, pages 3–14, ISSN 1613-0073, online `http://ceur-ws.org/Vol-2695/paper1.pdf`, Aachen, June 2020.

VII Minna Tamper, Petri Leskinen, Eero Hyvönen, Risto Valjus and Kirsi Keravuori. Analyzing Biography Collections Historiographically as Linked Data: Case National Biography of Finland. *Semantic Web Journal: Special Issue on Semantic Web for Cultural Heritage*, Mehwish Alam, Victor de Boer, Enrico Daga, Marieke van Erp, Eero Hyvönen and Albert Meroño-Peñuela (editors), Volume 14, 2, pages 385-419, IOS Press, ISSN 1570-0844 (P), DOI 10.3233/SW-222887, December 2022.

Author's Contribution

Publication I: "AATOS – a Configurable Tool for Automatic Annotation"

The DC (Doctoral Candidate) is the first author of this publication and wrote about 90% of the article. The DC transformed the Kansa Taisteli magazine articles to text to apply to them named entity linking. The DC also created the Kansa Taisteli magazine article perspective to the WarSampo portal in collaboration with A, B, D, F, and H (other authors based on the order in the article). C created the Semantic Finlex dataset in collaboration with D, F and H. The DC was the sole developer of the automatic annotation tool AATOS that used data and shared linking configurations developed in the WarSampo and Semantic Finlex projects with other authors. The DC applied the AATOS tool for named entity linking and subject indexing in WarSampo's Kansa Taisteli magazine article view and in Semantic Finlex contexts and evaluated the results.

Publication II: "Named Entity Linking in a Complex Domain: Case Second World War History"

The DC is the second author of this publication. The DC together with B and A devised the idea for the paper. The DC applied named entity linking methods to the Kansa Taisteli magazine texts using the AATOS tool. The DC evaluated the results of the AATOS tool and contributed to the planning and implementation of the evaluation for other datasets with A and B. The DC participated in creating the modular architecture for Named Entity Linking in the Second World War History domain with other authors. The DC wrote approx. 30% of the article, namely parts regarding own work.

Publication III: “Using Biographical Texts as Linked Data for Prosopographical Research and Applications”

The DC is the first author of this publication and wrote about 50% of the article. The idea for this paper was developed by the DC and C. The DC, A, and C devised the idea for transforming CSV data and texts into an RDF dataset. A designed the general model and transformed the CSV data into RDF and created the BiographySampo faceted search tool and data recommendations. The DC designed and implemented the transformation of the biographical texts into RDF. Based on the data, the DC got the idea for creating language analysis tools for BiographySampo. The DC designed and created linguistic application perspectives of the biographies. The idea of applying network analysis on the dataset was devised by the DC, A, and B. The DC participated in the design and implementation of visualizations for network analysis with Gephi by solely extracting the manual annotation links from the biographies and visualizing them with B.

Publication IV: “Finding Nineteenth-century Berry Spots: Recognizing and Linking Place Names in a Historical Newspaper Berry-picking Corpus”

The DC is the second author of this publication and wrote approx. 30% of the paper. The idea of the paper was proposed by A in collaboration with B and the DC. The research data was collected by A. The DC designed, implemented, and wrote about named entity linking for extracted place names. The disambiguation algorithm was proposed by the DC and was developped in collaboration with A. The entity linking and disambiguation was evaluated by A. The place names were extracted by B and evaluated by B and A for the task.

Publication V: “Visualizing and Analyzing Networks of Named Entities in Biographical Dictionaries for Digital Humanities Research”

The DC is the first author of the publication and wrote about 90%. The idea for this paper was devised by B and the DC. The DC devised the idea for named entity extraction and linking with B. The DC designed the ontology model for named entities, the ontology engineering process, and was in charge of the technical implementation of the linguistic knowledge graph. The named entity linking tool was created, executed, and evaluated by the DC. The network analysis view was implemented mainly by A into BiographySampo. Regarding the network analysis view, the DC contributed to the design by devising ideas about required toggles based on earlier experiences with Gephi. Based on an idea of explaining the links by B and the DC, the DC created the view for explaining the network references with sentences. The view was expanded with visualizations based on the DC’s idea about visualizing person’s historical significance that was also implemented by the DC. The contextual reader application was devised by B and the DC while the linguistic analysis perspective were also devised and created by the DC.

Publication VI: “Modeling and Publishing Finnish Person Names as a Linked Open Data Ontology”

The DC is the first author of this publication who also wrote about 80% of it. The idea for the paper was proposed by C and the DC. The idea for the ontology was from the DC and its use also in the presented services. The DC harvested and converted the Digital Agency dataset of names, acted as the main designer and developer of the HENKO – the person name ontology data model, ontology, and data service. The DC also created the person name finder service and transformed the Gender Identification system into a service that uses HENKO. The Gender identification system was originally created by A. A also transformed data from Sampo systems to be imported into the HENKO ontology by the DC. B managed the publication infrastructures for the tools and HENKO ontology.

Publication VII: “Analyzing Biography Collections Historiographically as Linked Data: Case National Biography of Finland”

The DC is the first author of this publication and wrote about 60% of the article. The idea for the article was proposed by B. The DC wrote most of the article and contributed to general data visualizations about the dataset, biographees, and vocations. The DC created statistics and visualizations related to authors, text analysis, and network’s reference analysis, while A for relatives, events, maps, and network metrics. The statistics and visualizations for references by gender and between relatives section was split evenly between the DC and A. All the visualizations were analyzed and evaluated in collaboration with D and C.

List of Tables

Abbreviations

3d Three-dimensional

API Application Programming Interface

BIBFRAME Bibliographic Framework Initiative

CH Cultural Heritage

CIDOC ICOM International Committee for Documentation

CIDOC CRM CIDOC Conceptual Reference Model

CRM *see CIDOC CRM*

DC Dublin Core

DCMI DC Metadata Initiative

DH Digital Humanities

EDM Europeana Data Model

FRBR Functional Requirements for Bibliographic Record

ICOM International Council of Museums

IE Information Extraction

IR Information Retrieval

HENKO Person Name Ontology (in Finnish Henkilönimiontologia)

LAF Linguistic Annotation Format

LOD Linked Open Data

NAF NLP Annotation Format

NBF National Biography of Finland

NE Named Entity

NED Named Entity Disambiguation

NEL Named Entity Linking

NL Natural Language

NER Named Entity Recognition

NIF NLP Interchange Format

NLP Natural Language Processing

OCR Optical Character Recognition

OWL Web Ontology Language

POS Part-of-Speech

RDA Resource Description and Access

RDF Resource Description Framework

SeCo Semantic Computing Research Group

SKS The Finnish Literature Society (in Finnish Suomalaisen Kirjallisuuden Seura)

SKOS Simple Knowledge Organization System

SPARQL SPARQL Protocol and RDF Query Language

TF-IDF Term Frequency – Inverse Document Frequency

TEI Text Encoding Initiative

UD Universal Dependencies

UI User Interface

YSO General Finnish Ontology (in Finnish Yleinen suomalainen ontologia)

1. Introduction

1.1 Background

In the recent years, Cultural Heritage (CH) [206, 239, 20, 214, 235] collections have been digitized to different databases. The CH collections can consist of artifacts, such as documents (e.g., letters, books, biographical descriptions). The collection data by nature is syntactically and semantically heterogeneous, multilingual, semantically rich, and highly interlinked. It is produced and hosted by organizations, such as museums, libraries, archives, and media organizations, and individuals. As the collections have been digitized, the role of information retrieval (IR) [152] has become relevant to the users of the data. It is often laborious to look for the data from multiple different databases and collections. In addition, there may be more information about the search topic in the database itself than what is evident based on search results. Documents can contain individual references to, e.g., people or places that can be hard to find with a search engine query. These references can contain new information and themes about these topics that can be lost in the results of search engines. The recall of traditional search engines can also suffer because they cannot handle, for example, people whose names change. Therefore, it can be laborious to study the depths of a database for breadcrumbs of information in order to get access to these references. Regardless, these document collections are studied in humanities research to learn about their content.

Digital Humanities (DH) [66, 204] is a field of research that utilizes computational methods to study digital resources and their usage. Computer science is the study of algorithms, computation, and information. The fundamental underlying question in computer science is, “what can be automated?” [5]. Recently, the application of computational methods, such as natural language processing [115] (NLP), has increased in DH research. The NLP supports application of automated methods for digitized texts, such as distant reading [168, 167, 109]. Distant reading applies computational methods to study texts, e.g., literature. The DH scholars use these methods to analyze digitized text collections and their content. In Finland, document collections, such as old newspapers, letters, and year books, are being digitized to improve accessibility and to preserve them for scholars. The NLP methods can be applied to these materials in order to facilitate data analysis. Similarly, NLP methods can be used to extract information that can be used to improve accessibility for search engines.

Semantic Web and linked data technologies can be utilized in describing digitized document collection metadata. The Semantic Web [14, 57] is an extension of the World Wide Web and its goal is to make data on the Web machine-readable. This can be achieved by adding descriptors to existing content and data on the Web by using technologies, such as the Resource Description Framework (RDF) [42]. Linked data [18, 13] is a term coined by Tim Berners-Lee to describe an interlinked web of data. The technologies enable description of text collections, linking them internally and externally to other document collections and other resources based on the metadata and used ontologies. Metadata is structured data that describes characteristics of an entity, such as a document [38, 178]. Ontologies are used to define the terminology between different agents and databases [216, 72, 23, 73]. In computer science, an ontology is defined as a shared model describing entities and their relationships in a specific domain [233, 108], i.e., a controlled vocabulary represented in a declarative formalism. Thus, ontologies contain concepts with explicitly defined semantics that can be used in metadata as values of metadata elements or properties. [72, 217] Describing document metadata with concepts of an ontology makes them understandable for machines [197, 10, 53, 31]. This enables, for example, the integration of heterogeneous document collections [110, 215].

Typically, Semantic Web content is published for machines by using linked data services through, e.g., SPARQL (SPARQL Protocol and RDF Query Language [227]) endpoints [80] functioning as an infrastructure that can be used to build linked data applications (e.g., web-based semantic portals and services) for end users. In recent years, the Semantic Web and linked data technologies have moved towards describing texts document structures to enable a variety of tasks, such as question answering [210, 135], describing documents (e.g., subject indexing) [221], and content recommending [162].

The Semantic Web and linked data technologies can be used by DH scholars to study digital collections and their application in the humanities research. The DH scholars rely on NLP tools and methods to process natural language or semi-structured texts to be able to apply data analysis to them. In order to produce data for close and distant reading applications, such as network visualizations, the data must often undergo various transformations and processes. By preprocessing text into a machine-readable format and saving it into a data storage for further processing, the data can be used by DH scholars to study and to apply data analytic methods to it. Using Semantic Web technologies not only makes the data machine-readable, but also enables describing the transformed text documents and their content to improve interoperability and accessibility. Interlinking documents to each other based on content facilitates finding mutual references from the collection.

1.2 Research Environment

The research contained in this thesis has been conducted in SeCo as part of the WarSampo: Finnish World War II on the Semantic Web (WarSampo)[1], Semantic Web Publications – Texts as Data Services (Severi)[2], and Automatic Anonymization and Annotation of Legal Documents (Anoppi)[3] projects. The WarSampo [94] project's goal was to integrate, enrich, and publish Finnish WW2 data as Linked Open Data (LOD) using Semantic Web technologies, linked data, and NLP methods. The Severi project's aim was to develop automatic annotation technology and tools for transforming texts into linked data services and applications, such as the BiographySampo[4] [97] system. In the Anoppi project, the focus was to develop open automatic tools for semantic content description and annotation of documents. The tools can be used to extract information, such as references to person names, to enable anonymization of legal documents published as linked data to improve transparency of the Finnish legal system. In parallel, the tools can be used for enrichment of document metadata to support creation of intelligent applications based on linked data.

1.3 Objectives and Scope

The aim of this thesis is to develop linked-data-based methods and tools to transform natural language texts to knowledge for data enrichment, data search and exploration, and data analysis. As a solution, a knowledge extraction toolkit for DH is presented. It enables text analysis, such as applying close and distant reading methods, to digitized historical document collections. Combining this with Semantic Web technologies, the collections can be made more accessible for the DH scholars and the public. The data models, tools, resources, and methods introduced in this thesis

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

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

[3] `https://seco.cs.aalto.fi/projects/anoppi/`

[4] Available at `https://www.biografiasampo.fi/`; visit project homepage `https://seco.cs.aalto.fi/projects/biografiasampo/en/`, to learn more.

are built for processing Finnish language text documents. By utilizing the toolkit, text documents are modeled and transformed into linked data, creating a knowledge graph [180]. The knowledge graphs of document collections can be enriched by using the knowledge graph as source data for the NLP-based enrichment methods in the kit. The methods can be applied to it to support creation of linked data infrastructures that can be used to build search, exploration, and data analytical applications for humanists.

The objectives for the knowledge extraction tools for DH presented in this thesis are:

- **Model for text document collections (OBJ 1)** Provide a data model that is usable for search, exploration, and data analytical applications. The data model provides a framework for describing document collections by standardizing metadata classes, properties, and their values.

- **Pipeline for transforming text to linked data (OBJ 2)** Provide a pipeline for transforming natural language texts to linked-data-based knowledge graph while preserving the text document's features, such as preexisting annotations and document structures.

- **Facilitate knowledge discovery (OBJ 3)** Support knowledge discovery [147, 185] by providing tools for enriching the knowledge graphs with, for example, referenced proper names, keywords, and linguistic information using the knowledge graph as a foundation for applications.

- **Applications for search, exploration, and data analytics (OBJ 4)** Provide pilot systems for searching, exploring, and analyzing document collections. The systems for exploring and data analytics also support close and distant reading.

- **Promote generalizability of tools and applications to other CH collections (OBJ 5)** Provide generalized tools and applications that can be used for Finnish language texts in different environments and text document collections.

- **Facilitate linked-data-based biographical and prosopographical research (OBJ 6)** Support scholars in biographical [199] and prosopographical research [240, 76] by enabling text analysis applications that apply close and distant reading methods on text document collections transformed into linked data. This Linked Open Data service can be used for querying, analyzing, and visualizing the data flexibly by using external tools, such as Yasgui [196] for SPARQL, or Jupyter[5] and Google Colaboratory (CoLab)[6] by Python scripting.

- **Evaluate the applicability of the data in proof-of-concept systems (OBJ 7)** Test the applicability of the knowledge graphs and its systems in practice by evaluating the datasets and proof-of-concept systems and analytics based on the data.

1.4 Research Questions

In respect to the objectives presented above, the goal of this thesis is to answer the following research questions.

1. **Can a data model using existing models (CIDOC, NIF, DC-Terms) model text document structure of the cultural heritage domain? (RQ 1)** For this research question, it is assumed that the target of modeling is a collection of natural language text documents. The data model for text documents of a collection needs not only to describe the collection but also to enable building applications. The model needs to include not only extracted information from the document collection, such as text structures, but also support adding enriched information, such as referenced proper names that are linked to internal or external source. These features are can be used for building search, exploration, and data analytical applications.

2. **The production of semantic data from text** Machine-readable data can be used to build applications for search, exploration, and data analytics that facilitate knowledge discovery. To produce the data from a document collection, a pipeline has to be built that first transforms and then enriches the data with, for example, entity linking. The production of semantic data from text is addressed by two research questions.

   (a) **Can a pipeline be built to transform Finnish documents and their collections be transformed into semantically rich and machine readable format? (RQ 2a)** This research question utilizes the data model from the previous research question for representing a natural language text document collection. The transformation requires attention to minimize the loss of information and to transform not only the texts but also add layers of semantic metadata, e.g., document structures and embedded annotations, into the machine “understandable"

[5] https://jupyter.org/

[6] https://colab.research.google.com/notebooks/intro.ipynb#recent=true

format. The NLP pipeline needs to be generalized so that it can be used for different domains and collections. In addition, it would be desirable if the pipeline could be applicable to documents in other languages.

(b) **Can NLP-based methods can be used to build generilizable tools to enrich knowledge graphs for different document collections? (RQ 2b)** This research question assumes that there is a knowledge graph containing document collection metadata. The documents can be used to enrich the metadata by extracting information and linking it to ontologies to bring added value, e.g., interoperability, to the knowledge graph. Therefore, by building tools to enrich the dataset to facilitate knowledge discovery. The entity linking tools require configurations that can change with respect to context. The tools should also be generalizable to be usable for different document collections.

3. **Can the semantic data enriched with linked named entities, keywords, and linguistic data be utilized to build search, exploration, and data analytic tools and applications for prosopographical research? (RQ 3)** This research question assumes that there is a document collection knowledge graph that contains enriched document metadata. The linked data infrastructure supports building applications on top of knowledge graphs. The knowledge graphs need to contain content describing metadata that is interlinked with ontologies. They can be used to build applications for search, exploration, and data analytics. The close and distant reading applications utilize the metadata to visualize the content. The applications built here are required to be generalizable and applicable to various document collections.

4. **Can semantic data and its applications enable new data analysis methods in biographical and prosopographical research? (RQ 4)** In this research question, it is assumed that a text document collection has been transformed in accordance to the data model of the first research question into a knowledge graph. It is also assumed that the knowledge graph is enriched and can be used through linked data services to build applications. These applications enable to study document collections and their provenance. The linked data services of a document collection knowledge graph enable querying and visualizing data directly for close and distant reading methods. This should facilitate biographical and prosopographical research.

In Table 1.1, the presented research questions (RQ) are answered with publications (P) I–VII. The table illustrates how each publication contributes to research questions.

Table 1.1. The relationship between research questions (RQ 1–5) and publications (P I–VII)

Research Question	PI	PII	PIII	PIV	PV	PVI	PVII
RQ 1	-	-	x	-	x	-	-
RQ 2a	-	-	x	-	x	-	x
RQ 2b	x	x	x	x	x	x	x
RQ 3	x	-	x	-	x	-	-
RQ 4	-	-	-	-	-	-	x

1.5 Research Process and Dissertation Structure

The research conducted in this thesis has utilized the design science [154, 84, 181] and action research [12, 11, 28, 44] methodologies. The design science is a utility and technology-oriented research paradigm in information systems discipline. Its purpose is to "devise artifacts" to attain its goals [84, 154]. These artifacts are assessed against criteria of value or utility to see if they work or provide improvements to existing solutions. Typically, design science research does not produce theoretical knowledge but applies knowledge of tasks or situations in order to create effective artifacts. These artifacts can be constructs, models, methods, and implementations. The process to create these artifacts consists of iteration of two main tasks, building and evaluating.

Design sciences are typically applied sciences. In the building task, the artifact or new technology is constructed for a specific purpose by exploiting the knowledge created by basic research. However, it can prove to be a challenge to explain how and why the created artifact works as the natural laws governing an artifact and the surrounding environment are not well understood. Therefore, the knowledge and understanding of a design problem and its solution are gained while building and applying an artifact. The novelty of design science, in comparison to routine design work and systems building work, is that its purpose is to create an innovative, purposeful artifact for a specified problem domain. This means that the solution is for an unsolved problem or a known problem is solved in a more effective or efficient way. In design science, the progress is made when existing technologies are replaced by more effective ones. The evaluation task determines the performance of the product based on its purpose. The applicability of the solution is evaluated using experimental, analytical, observational, testing, and descriptive evaluation methods in a real world environment and scenarios.

In contrast to the technology-oriented design science, action research is a research method that is considered suited to the study of technology in its human context [12, 11, 44]. Action research's focus is on solving organizational problems with subjects of the research while at the same time contributing to knowledge. It is a process that consists of two main stages: a diagnostic and a therapeutic stage. In the diagnostic stage, the researcher and the subject collaborate to get an analysis of the social situation. This is followed by the therapeutic stage where changes are made to improve the situation and their effects are studied.

The action research is by its nature an iterative process and it has additional structure to achieve required scientific rigor. After establishing a research environment, the action research cycle iterates five identifiable phases: diagnosing, action planning, action taking, evaluating, and specifying learning. The evaluation phase must analyze whether the taken action relieved the problems. Regardless of the outcome of the action, the action research cycle can continue and adjust its theoretical framework to reflect the outcome. Action research aims to link theory and practice by developing further knowledge about the organization and the validity of relevant theoretical frameworks until it has managed to understand or solve a given problem.

In this thesis, a knowledge extraction toolkit is created that contains models, tools, and methods for analyzing text document collections. The toolkit is based on analyzing the user requirements that are collected from scholars, professionals, and laymen. The user requirements have been collected by conducting literary and systems reviews, formulating illustrative use scenarios, building pilot prototypes based on the data, and gathering feedback from the actual users, such as scholars in humanities. Based on the iterative nature of the design science and action research methods, mathematical and user-based evaluations have been used to assess the accuracy and the purposefulness of the developed tools, transformation of the data, and the data analytical applications. The prototype tools and applications act as proof of concepts that demonstrate the utility of the software and data artifacts for the given user requirements [181]. In addition, by using the system and the data in real world use cases in the action research setting evaluates its impact in actual situations. The utility of the systems is ensured by basing the features of the data, functionalities of the tools and applications on existing research, and illustrative use scenarios.

This thesis is structured as follows. The Chapter 2 contains the theoretical foundation and literature review. The results of this thesis are presented in Chapter 3. Lastly, the implications of the results, validity of the work, and further research are discussed in Chapter 4.

2. Theoretical Foundation

In this chapter, the theoretical foundation of the research questions and objectives of this thesis are reviewed. The literature review is divided into modeling document collections, automatic annotation pipeline implementations, and applications of the data. In Chapter 4, the comparison between literary review and results are discussed in detail.

2.1 Modeling Document Collections

The CH collections consist often of multi-disciplinary data (archaeology, history, architecture, science, etc.) that are available in multiple formats (images, texts, 3d models, etc.) and aimed to different user groups (museum operators, tourists, academics, etc.) [235]. To help different user groups to find this data, various collective data aggregation systems have been created and opened for the public to browse the data [235]. These data aggregation systems exist not only on international, and national level, such as Wikidata[1][244], Europeana[2] or Finna[3] but also in thematic communities, such as ARIADNEplus[4] in archaeology.

Modeling these collections has proved to be a challenge from an IR perspective, and Semantic Web technologies have been proposed to solve the issues [235]. There are standardized vocabularies and models that have been designed for representation of different structured controlled vocabularies, such as RDF Schema [245], Simple Knowledge Organization System (SKOS) [165, 107, 235] and OWL (Web Ontology Language) [87]. Several data models have been created also for describing features of documents, their collections, and related metadata, e.g., document name, publication time, collection maintainer, and location of the collection. The Europeana Data Model (EDM) [50, 235] is a data model designed for de-

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

[2] https://www.europeana.eu/

[3] https://www.finna.fi/

[4] https://ariadne-infrastructure.eu/

scribing CH collections. There are also specified models and vocabularies for bibliographical cataloging, e.g., Resource Description and Access (RDA) vocabularies in RDA Registry [229, 184], Bibliographic Framework Initiative (BIBFRAME) vocabulary and model [131, 249], and Functional Requirements for Bibliographic Record (FRBR) [228, 235]. For describing resources available online, there are general vocabularies, such as Schema.org [74], Dublin Core (DC) Metadata Element Set [46, 235], and DCMI[5] Metadata Terms [190, 45, 235]. In addition, the International Committee for Documentation (CIDOC) has developed since 1996 a Conceptual Reference Model [49, 51, 235] CIDOC CRM that has been registered as a ISO standard and aims to facilitate information interchange between databases of museums, libraries, archives, and others. These models are utilized in data aggregation systems that use Semantic Web technologies and linked data, such as in Europeana, and ARIADNEplus.

The document collections consisting of artifacts, such as newspapers, biographical documents, or books, can be part of library, museum, or archive collections. For example, the Finnish National Library[6] has digitized a vast collection of books, pamphlets, newspapers, and other textual documents for users. The texts can be searched from databases based on the metadata that has been collected from its items. The metadata of these databases often consists of elements, such as the title, authors, publication year, and publisher. Content has often been described using keywords. Some organizations and institutions, such as libraries [203], have moved to using Semantic Web and linked data in describing their document collections. The National Library of Finland has published as LOD the national bibliography Fennica [222][7], containing the largest collection of bibliographical entries in Finland.

Modeling texts of document collections has also received more attention in the past few years. Well-known standards and data models for modeling texts include Text Encoding Initiative (TEI) [238, 225], NLP Interchange format (NIF) [81, 82, 83], and NLP Annotation Format (NAF) [60]. The TEI is a standard for the representation of texts in XML format. Unlike the RDF data model and language, the XML format does not provide a computational semantics to markup. TEI format is not by default compatible with the RDF and needs to be converted to an RDF-based format, such as RDF/XML, to achieve that [37, 113]. The NIF is an RDF/OWL-based format and its goal is to attain interoperability between NLP tools, language resources, and annotations [83, 155]. It concentrates on modeling document structures, common terms, and concepts. The NIF model has been utilized in building tools [121], RDF-based corpora for machine learning [26], and in NLP tasks, such as question answering [210, 135] and

[5] Dublin Core Metadata Initiative

[6] https://digi.kansalliskirjasto.fi/

[7] https://www.kansalliskirjasto.fi/en/news/finnish-national-bibliography-released-open-data/

a transformer has been created for converting CoNLL-u to NIF [33]. In contrast, the NAF is originally an XML-based model designed to represent linguistic annotations in complex NLP architectures. It complies with the recommendations set for the XML data model created for modeling language called Linguistic Annotation Framework (LAF) [102, 103]. The NAF model can also be converted to RDF and, similarly to the NIF, its representations conform to the principles of linked data. The NAF can also be used to describe named entities (NE), e.g., referenced names that are linked to external sources within the text. It has been used to model, for example, the BiographyNet document collection and news of the NewsReader dataset [61, 60]. However, the NAF would require more adjusting to describe document structures and their relations in more detail whereas the NIF is more widely used but lacks the ability to model language and linguistic information. Also, the NIF enables the modeling of NEs extracted from the text, however, similarly to the NAF, it is a generic representation of a NE that doesn't take into account the different purposes of the possible co-existing annotations, such as manual annotations (e.g., HTML links, highlighted concepts or works) and automatically generated annotations.

There are multiple formats that can be used for modeling language and linguistic information. The CoNLL-U format[8] is one such an example. The format has been created in the Universal dependencies (UD)[9] project [173] as a standard for annotating linguistic information, such as part-of-speech (POS) tags, lemmas, dependency grammar, and structures. Also, RDF-based vocabularies [159] have been developed for more complex linguistic data, such as Lemon [157], OntoLex-Lemon [158], and the Multilingual Morpheme Ontology (MMoOn) [123]. The model for the MMoOn has influenced the OntoLex-Lemon model development [34, 124]. The NAF format utilizes guidelines of the XML-based LAF model in annotating linguistic information. The LAF data model was developed by the International Standards Organization (ISO)'s sub-committee on Language Resource Management [103]. The LAF's purpose is to provide a broad framework for representing linguistic annotations to served as the basis for lexicons, morpho-syntactic and syntactic annotations, as well as for a range of semantic annotation types. The LAF-based linguistic annotations have been revised in the NAF format [60] in order that it can be easily be converted to RDF. It supports several annotations over a text at different linguistic levels including information, such as the POS tags and lemmas of terms. However, it would need to be adjusted to support morphological features of Finnish while the CoNLL-U format, for instance, would provide already grounds for annotating linguistic information in Finnish language texts.

[8] https://universaldependencies.org/format.html

[9] https://universaldependencies.org/

2.2 Production of Semantic Data from Text

Manually annotating and subject indexing document collections is laborious, costly, and time consuming work. [36, 137] However, it can also be a hard task to do annotations automatically with a computer and it requires dedicated algorithms and domain specific information for the task. Automatic annotation process utilizes NLP methods for information extraction (IE) [40] to pick information, such as POS tags for words, events, person names, place names, topics, or keywords, from unstructured or semi-structured text in natural language documents. The annotations can be added into document metadata and used, for example, in search optimization in search engines [52, 38, 183, 128] or in different data visualization and analytical applications (e.g., network [55, 247] and map visualizations [86]).

In Finnish language materials, IE methods, such as subject indexing, have been used for documents, e.g., newspapers [120, 128], member or student catalogs [138, 95], and other documents [118, 117]. The methods are often used to make search applications to find documents based on extracted keywords, words, and mentioned names. Typically, these systems, depending on the source material, perform OCR on the texts, apply NLP methods, and finally use the results to build search applications and data analytical visualizations.

In this thesis, the automatic annotation methods are used in a pipeline that transforms text firstly to RDF and then to enrich its metadata for search optimization and data analytics. The NLP methods used are discussed in the following subsections. Automatic NLP pipelines and toolkits have been created for different languages earlier [121, 182, 136, 6, 194] but not all of them use RDF to record the results. The toolkits are often for users, such as scholars or software developers, and they contain means for transforming and annotating text into a desired format that can be then used for tasks, such as IR, data analytics, or information visualizations.

2.2.1 Extracting Linguistic Information

In order to extract linguistic information from Finnish language materials, the first stage is to extract linguistic and morphological information from the texts. Like Estonian and Hungarian, the Finnish language is highly inflected [220, 211]. In Finnish, meaning is expressed through morphological affixation (agglutinativity) by using case endings [220, 211][77, Section § 53]. The slightly inflected languages, such as English, use syntactic structures to express plural and possessive relations, grammatical cases, and verb tenses and aspects [220].

A common method for finding the root form of a word is stemming [152]. Stemming can be characterized as a crude heuristic process that cuts off the ends of words to find the root. The process often includes the removal of derivational affixes. Regardless, stemmers (e.g., the Snowball stemmer[10] [129]) don't work well with highly inflected languages [211, 119, 220, 3]. An alternative method for finding the root or base form of the word is to utilize lemmatization [153, 152]. Lemmatization's aim is to uncover the lemma (e.g., basic form or base form) for an inflected surface form of a word that appears in the input text [153, 152]. The lemmatization process utilizes vocabularies and morphological analysis of words to find their lemmas [153, 152]. Therefore, it is considered to be a more sophisticated method for finding the base form for a word than stemming.

The morphological analysis can be also used as a standalone method to find word's base form and to get linguistic information. Its goal is to identify the word's inflection form and use the information to find the base form. The morphological analysis selects the most probable sequence of lexemes [250, 4]. A lexeme is a unit of lexical meaning that includes a lemma and inflected forms [153]. There exist specially designed morphological analyzers [130] for Finnish language texts to tackle the complexity of the morphological structure of words, such as FinTwol[11] [144], OMorFi[12] [187], and FDG[13] [224]. They are required to uncover lemmas embedded in words [143]. In addition, in recent years many other tools have been created that use morphological analysis tools to enrich its results with linguistic information, such as the Finnish dependency parser[14] [79] and its follower Turku neural parser pipeline[15] [116]. They enrich the morphological analysis results with dependency grammar and transform the results to CoNNL-U[16] [193] format.

2.2.2 Named Entity Linking

The process of extracting and linking textual entities, such as place names, is called Named Entity Linking (NEL) [75, 27, 155]. NEL extracts and links NEs to vocabularies and ontologies from text input. The NEs are defined as proper names and quantities of interest [230, 35]. They are phrases that contain names of persons, organizations, locations, dates, times, and quantities (monetary values, percentages). According to Jurafsky [115], a NE is everything that can be referred to with a proper name. Named Entity Recognition (NER) is that task that identifies these entities from texts by detecting and classifying names from the text to different categories [230,

[10] http://snowball.tartarus.org/

[11] http://www2.lingsoft.fi/cgi-bin/fintwol/

[12] https://github.com/flammie/omorfi/

[13] http://www.connexor.com/

[14] http://turkunlp.github.io/Finnish-dep-parser/

[15] http://turkunlp.org/Turku-neural-parser-pipeline/

[16] http://universaldependencies.org/format.html

170, 70].

Accoding to Hachey et al. [75], NEL can be divided into three phases: *extraction*, *searching*, and *disambiguation* that produce the linked NEs. The NER task extracts NEs from texts. In the second phase, the descriptive identities are searched for the entities. NEL requires the use of vocabularies (such as Wikidata or Wikipedia (DBpedia) [205, 75]) that contain vast amounts of information about entities in order to be able to identify and connect to them. Named entity disambiguation (NED) [75, 27, 248, 41] is the process of signing correct identities or links for NEs. For example, the word *Suomi* can point to, for example, the last name of a person or a place name depending on the context. The use of context is vital in identifying the correct identities of NEs [88]. In Finnish, person names can be hard to separate from organization or place names [188] in addition to having multiple different persons or places with the same name. For example, *Karjala* can refer to a village or geographical area in eastern Fennoscandia. There are multiple proposed methods for general disambiguation of NEs and also for dedicated methods of particular NE types, e.g., geographical named entity disambiguation [71, 89, 65, 242, 29, 88, 253].

NEL tools [163, 172, 156] have been developed for different types of English language texts and use cases. DBpedia Spotlight [163] has been created to link text documents to DBpedia [8] to enable building search and faceted browsing applications. It uses the vector space model for NED. The DBpedia Spotlight can be configured for a variety of languages, however, the results are poor for Finnish. The Tagme [58] system, similarly to DBpedia Spotlight, has been developed to link text to DBpedia. It applies the bag-of-words paradigm [115, 78] that is enhanced with a voting scheme for NED for English texts. Other state-of-the-art tools for NEL include AIDA [172] (mapping algorithm), YODIE [43, 47] (NED based on combination of four metrics: textual similarity, semantic similarity between neighboring entities, word level contextual similarity, and URI frequency in Wikipedia articles), and Zemanta [47] (machine learning). These tools have been used, e.g., for analyzing and linking mentions in Twitter feeds.

For Finnish texts, the ARPA tool [148] has been developed. ARPA is an entity linking tool that can be configured to match entities from text to different vocabularies. It uses the Language Analysis Service (LAS) [149] for lemmatizing. In order to identify NEs, various NER tools can be used depending on the language. For english, there exists a vast variety of NER tools [236], such as Stanza [194], spaCy[17][236] and BERT-based NER models [48, 142]. The most commonly used NER tools for Finnish are Stanford NER[18][59], FinBERT-based Finnish NER [145], and FiNER [201]. Stanford NER is a machine-learning-based named entity recognizer tool that

[17] https://spacy.io

[18] https://nlp.stanford.edu/software/CRF-NER.html

can be trained for various languages including Finnish [202]. FinBERT is based on Google's BERT model and can be used for NLP tasks, such as NER and text classification. FinBERT is developed by the TurkuNLP research group. The FiNER tool is a rule-based tool that utilizes vocabularies in NER. It has been developed by the FIN-CLARIN consortium at the University of Helsinki.

2.2.3 Subject Indexing

Subject indexing [220, 36, 137, 252] is a process of describing the contents of documents using descriptive keywords from a controlled vocabulary or thesaurus. It aims to cover the main topics exhaustively and describe the aboutness of a text precisely, while seeking a condensed representation of the content. Many document collections use subject indexing and keywords to describe the document. Traditionally subject indexing requires that a human reads the document and then creates annotations or picks keywords. A generic tool for automatic subject indexing and annotating could reduce time and resources spent on the task [36, 137, 252].

Automated subject indexing tools [198, 189] use a process that can be divided into two main phases: keyword extraction and keyword assignment. Keyword extraction is a sub-task of IE that identifies the keywords from the given input text. The task utilizes typically a dictionary of extraction patterns that can be produced manually or created automatically. Usually systems that create automatically extraction patterns use resources for training, such as previously annotated texts that contain context specific markup or manually assigned keywords. The keyword assignment is the identification of keywords from a controlled vocabulary of a reference list (a thesaurus) that often contains semantic information and relations (semantic lexicon). The choice of the keywords is made based on the selected term weighting scheme. Typically, models use three main variables to evaluate the relevancy of a word in a document: term frequency, document length, and the rarity of the word in the document collection. These features are used together to evaluate how distinctive a word in a document is. The documents in a collection are rewarded for term specificity if they contain rare words [177]. A frequently used term weighting scheme is the TF-IDF (term frequency and inverse document frequency) [177, 152] where weight of an individual term is computed using the TF-IDF scores to determine its uniqueness in a corpus. Unique terms are considered more important than frequently mentioned terms.

Subject indexing has been applied to various digitized datasets. There are multiple tools and algorithms for automatic subject indexing [146, 7, 221]. Most tools for automatic subject indexing are either lexical or machine-learning-based approaches [146, 7, 221]. Applications using lexical approaches include Maui and KEA [161] while for machine-learning-based applications FastXML [191], PD-Sparse [251], and fastText [114] are a couple examples. There have also been attempts to create ensembles and/or fusion architectures utilizing both approaches. An example of this is the Finnish Annif [221] tool created at the National Library of Finland.

2.3 Explorative Applications and Data Analysis

In order to use document collections for IR and analytical purposes, different applications can be built. In digital humanities case studies it is often the case that the data is ready for use after it is first filtered and transformed into a suitable format (e.g., TEI, HTML, or RDF) and then imported into a information visualization tool, such as ArcGIS[19], Recogito[20] [207, 208], Palladio[21] [54], Gephi[22] [32, 160], or tailored using programming languages, e.g., R or Python [213, 86, 122, 246, 234, 90, 85]. These tools are then configured to visualize results to answer different research questions and study the dataset at hand. In some cases public demonstrators are built and made available online for wider audiences [195].

The search, exploration, visualization, and data analytical applications can be also used for various DH data collections, such as biographical data. Representing and analyzing biographical data has become a new field of research and application. The first Biographical Data in Digital World workshop was held in 2015 (BD2015), where multiple works were presented about studying and analyzing biographies as data [1], and followed by a new proceedings of BD2017 with similar works [2]. In [134, 166], U.S. Legislator registry data[23] was used to create search, exploration, and data analytical visualizations. In the NewsReader project[24] the goal was to create story arcs by extracting events and actors from unstructured news and transforming them into RDF [200]. As a part of the NewsReader project, language technology was applied to Dutch biographies in the BiographyNet[25] for extracting entities and relations and transforming them to RDF [174].

The Semantic Web and linked data infrastructure can be utilized to provide search, exploration, and data analytical services and applications for scholars. The semantic portals and services based on knowledge graphs usually provide the end users with means for data exploration [104] and

[19] https://www.arcgis.com/index.html

[20] https://recogito.pelagios.org/

[21] https://hdlab.stanford.edu/palladio/

[22] https://gephi.org/

[23] https://github.com/unitedstates/congress-legislators/

[24] http://www.newsreader-project.eu/

[25] http://www.biographynet.nl/

analysis. These portals and services provide users with variety of applications, such as faceted, ontology-based, and entity-based search engines, semantic browsing based on relations extracted and reasoned from the data, and tools for data analysis, visualization, and serendipitous knowledge discovery [92, 9]. An approach to provide these services is by utilizing the Sampo model[26] [93] that has been used in a series of semantic CH portals, such as BiographySampo [98], AcademySampo [139], WarSampo [101], and NameSampo [105]. The model's idea is that a semantic portal provides the end users with multiple application views to the content. The application perspective usage can be split to two phases. First, selecting interesting data using ontology-based faceted search [232, 127, 106]. Secondly, the visualization and data analytical tools are applied to the selected data.

In this thesis, the focus is on applications for network and linguistics analysis and how they can be built to study the modeled and enriched data extracted from natural language documents.

2.3.1 Network Analysis

Network science [22] studies complex networks, e.g., telecommunication networks, computer networks, biological networks, and social networks, where distinct actors or other entities are represented by nodes and the relations between them as links. It applies theory and methods from various fields of science, such as graph theory (mathematics), statistical mechanics (physics), data mining, and information visualization (computer science). A social network [176] is a social structure formed from a set of social actors (e.g., people or organizations), sets of dyadic ties (e.g., family relations or links to interests and works), and other social interactions between actors. In this thesis, we will concentrate on social networks formed from mentioned people extracted from a document collection.

Social networks have been used in information sciences [176, 141] and digital humanities [68] as a distant reading method. The network analysis based on biographical data has been studied in [247, 133, 25] where networks were created by extracting NEs and their relations from text. Several related works [247, 179, 133, 25, 55] about network analysis and visualization methods [171] have been presented for different datasets. The networks have been constructed from various formats and from datasets with sufficient metadata. Some networks [247, 179] enable browsing of the underlying data and relationships between its objects. In these networks, the variety of toggles (e.g., coloring and weighting options for nodes and links based on certain data derived criteria) for adjusting the network are limited and the networks offer a controlled set of nodes that can be browsed to study the underlying collection metadata.

Such exploratory relationship analysis views are also often limited in the

[26] `https://seco.cs.aalto.fi/applications/sampo/`

network visualization applications. In the case of LinkedJazz[27], there is a relationship view that gives more insight and helps to understand the links between some people of the network. The view shows transcripts where the references are made to other jazz musicians.

2.3.2 Linguistic Analysis

The corpus linguistic analysis [69] or just linguistic analysis is a popular method when studying the language use in different materials and document collections among linguists. Linguistic analysis applies corpus linguistic methods to a corpus or dictionary of annotated natural language texts. Corpus linguistics [112, 17, 231] is a research approach that uses quantitative or qualitative methods for obtaining and analyzing language data from a natural language text corpus. It, however, goes beyond methodological role and allows researchers to ask fundamentally different kinds of research questions, that sometimes results in changes in observations that derive from this approach [17, 231]. It is a discipline within humanities that has utilized computational methodologies, such as NLP [24, 56], and is a branch of digital humanities research.

Linguistic analysis has been used, for example, to study parliamentary speeches [19]. In addition, linguistic analysis has been used in other fields, such as in Ko et al. [126] where a study was made about the computer software bug reports, with an analysis of the usage of verbs, adjectives, conjunctions, prepositions, and adverbs. In digital humanities, digital methods have been applied to create tools and portals where the user can study the vocabularies. One notable tool for digital humanists is the Voyant Tools[28] [209] that supports many languages and their text analysis. For Finnish language resources, there is the KORP[29] [21] tool. The KORP tool contains multiple corpora from Finnish newspapers to internet forum discussions. The user can search for words in the selected corpora and get a list of sentences that mention the word, statistics for corpora, and a geographical map where the words are used based on the location metadata about the document context.

Applications for visualization of NEs in texts have been created before. The purpose has been usually to interlink document collections, such as Wikipedia, based on mentioned NEs or terminology to provide more context and information to the users. For example, the contextual reader application, CORE [151], was created to find NEs in real-time from documents, link them to configured ontologies, and visualize the results by replacing the mentioned NEs in the texts with links that provide context. Similar vi-

[27] https://linkedjazz.org/network/

[28] https://voyant-tools.org/

[29] https://korp.csc.fi/

sualizations of NE data have been used in, e.g., DBpedia Spotlight[30] [163] and Gate Cloud[31] [156].

[30] https://www.dbpedia-spotlight.org/demo/

[31] https://cloud.gate.ac.uk/

3. Results

This chapter answers the research questions of this thesis by presenting the results from the publications of this thesis. Afterwards, the results presented here are compared to prior research in the Chapter 4.

3.1 Modeling Document Collections

The data model for text documents can influence its usability for different tasks. This section describes the data model that is developed to answer to the research question 1: Can a data model using existing models (CIDOC, NIF, DC-Terms) model text document structure of the cultural heritage domain? The main research objectives of this section is the OBJ 1: Model for text document collections. The contributions are presented in the publications III and V that describe how to model a biographical document collection.

The text document collection modeled in this thesis originates from the National Biography of Finland [1] (NBF), edited by the Finnish Literature Society (SKS, in Finnish Suomalaisen Kirjallisuuden Seura) and published as a series of ten volumes [125] in print in 2003–2007. The collection is supplemented with other biographical collections published by the Finnish Literature Society, e.g., the Finnish Clergy 1554–1721 and 1800–1920, the Finnish Generals and Admirals in the Russian armed forces 1809–1917, and the Finnish Business Leaders, totaling today over 13 100 biographies [99]. The collections consist of a collection of short biographical texts written by the experts in the fields of history, science, art, culture, and business. They have also been made available online[2]. The collections were re-published in 2018 as the semantic portal *BiographySampo—Finnish biographies on the Semantic Web* [97].

The linguistic data model, described in Publication III, for BiographySampo's biographical text document collections was created by utilizing

[1] https://kansallisbiografia.fi/

[2] https://kansallisbiografia.fi/english/

generalized models, such as NIF, DC Terms, and CIDOC CRM. It is created to enable building search, exploration, and data analytical applications for the dataset. The use of these data models facilitate information interchange between other document collections in services hosted by institutions, such as museums, libraries, and archives that use the same data models. The CIDOC CRM is used in the model by adding the *cidoc:E31_Document* class for each document object. The NIF model is used to separate document structures, such as titles, paragraphs, sentences, and words. The NIF model also supplied the document objects with text representations for paragraphs and titles and properties for describing the relations between sentences and words. DC Terms is used for describing the relationships between text structures by adding relations, such as isPartOf or hasPart, between paragraphs, sentences, and the document. The CoNLL namespace, used for the RDF representation of the original CoNLL-U format [33], is used for describing results of morphological analysis results for words from POS tags to word dependencies. The use of the conll namespace enables transforming CoNLL-U format to RDF while minimizing the information loss in the conversion process. In addition to the existing metadata schemas, a custom vocabulary was created to supplement the model. The custom vocabulary includes properties and classes for modeling features, such as HTML annotations of the texts, their relationships with text structures, and ordering of text structures. The HTML annotations were divided to classes depending on their annotation tag name and purpose, e.g., link or emphasis. Annotations also have the original text as a label and a possible link. The instances are connected to the source paragraph objects using DC Terms' *dct:references* property and to words using *dct:hasPart* property.

The knowledge graph, based on this data model as such, would have been laborious to use. For example, some applications utilizing the data need to be able to separate the natural language texts from the semi-structured text before using it. Therefore, as a part of the solution to research question 1, a custom vocabulary for a group of helper properties and classes needed to be established. The helper properties supplement the model, for example, in classifying text based on content type (e.g., lead paragraph, text paragraph, or references). With the helper classes and properties, the analytics or NLP tasks based on the data, such as subject indexing, is easier to implement because, for example, the text paragraphs can be queried directly. Also, helper classes with statistical figures for each document (e.g., word counts, verb counts) can help in language analysis based on SPARQL by making the queries faster by reducing the amount of work.

In addition to modeling the texts, the knowledge graph was also enriched with automatically extracted and linked NEs (described in Publication V). The NEs were modeled using the custom vocabulary, NIF, and DC Terms.

The NE model differentiates the NEs from other types of annotations extracted from the texts. In the model, each entity is classified as a NE and they have basic information, such as extracted string, lemma, NE type, links to related concepts in other vocabularies, the location of the string in text, and used extraction method of the entity. The NE objects are also connected to the related words that are a part of the corresponding NE using *dct:isPartOf*. The sentence instances also are connected to NEs using a custom property. The extraction methods of NEs are collectively described in their own class that supports adding more provenance information for the method. Similarly, the NEs are typed using classes that describes the NE types. In addition to these, the overlapping NEs extracted from the text (due to multiple results from different tools) are grouped using a class representing a NE group. Each group makes a distinction regarding which is the primary entity that has the highest score among overlapping entities. The other entities of the group are simply the group's members.

The data model presented here describes the model created to represent the BiographySampo's biographical collection in RDF format. It has been created to provide easy access to the data for multiple different purposes, such as data analytics. The model also utilizes existing data models to facilitate information interchange between other document collections. The novelty of the model is that it supplies the text structures with detailed linguistic information (e.g., POS tagged words, dependency grammar) and differentiates between different types of annotations in the texts, e.g., HTML annotations and automatically extracted NEs. It also provides means to classify text paragraph types, e.g., lead, body text, and semi structured paragraphs. The data model offers a starting point for many NLP tasks, such as subject indexing and topic modeling, that can be used to enrich it further.

3.2 Production of Semantic Data from Text

The data model can be used in transformation of the text document collections to a knowledge graph that can be used for building applications and data analysis. The size of text document collections can vary a lot. Processing many documents can be a time consuming process depending on the size of the collection. Maintaining document structure is important to data analysis and NLP tasks that require context. Also, extracting pre-existing annotations can be useful in data analytical applications and in studying the document collection. Therefore, the procedure for transforming text into RDF using the model must make sure that the document structures and pre-existing markup are preserved. The extracted data can also be used in enriching the dataset further, e.g., with linguistic information or to identify references to NEs. To achieve this, an annotation pipeline is required for transforming document collections into RDF and enriching them. The annotation pipeline extracts from the text 1) document structures, 2) pre-existing markup, 3) linguistic information, and 4) content describing entities. Here, the production of semantic data is split into two parts: 1) transforming document collections into RDF and 2) enriching the document collection using NLP techniques. Thus, the research question 2 addressing this matter is split into two questions (2a and 2b). The solutions to research questions and objectives are explored separately in more detail in the following sections.

3.2.1 Text Transformation Pipeline

Text document collections transformed into RDF can be useful for building applications that utilize the knowledge graph. These applications require a knowledge graph that contains machine-readable semantically rich data. This section answers the research question 2a: Can a pipeline be built to transform Finnish documents and their collections be transformed into semantically rich and machine readable format? The main research objectives of this section are OBJ 2 (Pipeline for transforming text to linked data) and 5 (Promote generalizability of tools and applications to other CH collections). The publication III describes the transformation of the BiographySampo document collection into RDF.

An NLP pipeline was developed to enable transforming document collections into linked data. It produces RDF using the model described in Section 3.1. The document collections can be in different formats depending on the source data. The pipeline needs to support different source formats to enable transformation of texts in accordance with the model. Therefore, the pipeline can be configured to extract text from RDF, text, or HTML input. First, the pipeline creates a skeleton that contains metadata, such as relations and ordering, about the document structures, paragraphs, and titles. This is followed by a process where the pipeline utilizes existing software components, such as the Turku dependency parser, to generate morphological analysis of the document texts in CoNLL-U format. The CoNLL-RDF [33] tool then transforms the results into RDF. At the time, these tools were the only ones that are freely available, best performing, and least laborious to apply to transform text as RDF. These tools are supplemented with a reasoner module that infers more properties and connections based on the previous results and renders its results into RDF. For example, the Reasoner module adds order numbers for sentences within a paragraph to help to sort them and relations to connect sentences and words directly to documents. The Reasoner aims to enable easier queries with the added relations to retrieve sections of a document, such as sentences, in correct order.

The NLP pipeline can process documents that can contain HTML annotations. It was utilized to transform the biographies that are part of the biography collections. The collection's texts contain HTML markup, such as paragraph tags, links to other biographies, and emphasis tags highlighting works of art or quotes. The pipeline extracts these HTML annotations from the text and uses them to divide the text into paragraphs and titles. In addition, the HTML links and emphasis tags were added into the RDF representation. Afterwards, the annotations are removed from the text representation of the document sections that are added into the dataset.

The classification and transformation of text into RDF succeeded with 100% for paragraphs, 99.5% for sentences, and 99.0% for words (Publication VII, Section 2.2). During the process, the HTML links to other biographies were extracted from the text documents with accuracy of 99.4% and the links were added into the document metadata. The results presented here, were calculated for 200 randomly selected entities in each category (Publication VII, Section 2.2).

3.2.2 Enriching Document Metadata

Text document collection knowledge graphs hold huge amounts of data. It can be hard to search for documents of the knowledge graph without metadata describing the content. This section explores enriching of knowledge graphs with metadata and answers the research question 2b: Can NLP-based methods can be used to build generilizable tools to enrich knowledge graphs for different document collections? The main research objectives of this section are OBJ 3 (Facilitate knowledge discovery) and 5 (Promote generalizability of tools and applications to other CH collections). The results presented here are described in more detail in publications I, II, III, IV, V, VI, and VII. For enriching document collection metadata a number of tools have been created for enriching document metadata using named entity extraction, subject indexing, and entity linking. The tools utilize pre-existing libraries and software modules freely available, best performing, and least laborious to apply at the time. These tools and applications have been described in the following sections for each use-case.

A document collection knowledge graph can be enriched with various automatic annotation tools. The tools can be used to extract data, such as morphological information about words, NEs, and keywords, and link them to their corresponding ontologies. In this thesis, existing document collection knowledge graphs were enriched with content describing metadata in BiographySampo, WarSampo and Semantic Finlex systems. The WarSampo dataset [101] contains Kansa Taisteli magazine articles and their metadata. Kansa Taisteli magazine was published by the Sanoma Ltd and the Sotamuisto association between 1957 and 1986. [226] The metadata has been manually collected by Timo Hakala [226] and converted into an RDF format by Kasper Apajalahti. Semantic Finlex[3] [175, 91, 63] is a service that provides Finnish legislation and case law as LOD. The knowledge graph contains court decisions, statutes, and their metadata that has been transformed into linked data from the legal database Finlex Data Bank[4] of the Ministry of Justice. In the case of BiographySampo, the NLP pipeline was used to first create the dataset and then to enriched it similarly to the other knowledge graphs.

In Section 3.2.1 the creation of NLP pipeline for transforming the document collections was described. In addition to transforming the document collections, the NLP pipeline's data conversion process in BiographySampo included enriching of the original document collection data with linguistic information, such as results of lemmatization and morphological analysis. The morphological analysis results were used also in identification of text structures that are described and evaluated in Publications III and VII. The NLP pipeline recorded POS tags, lemmas, morphological features, and Dependency Grammar information into the model. The NLP pipeline uses the Turku dependency parser that attained accuracy of 95.6% for POS tags (Publication VII, Section 2.2). The result was calculated for 200 randomly selected word entities with POS tags. They were closely the same as reported in the original evaluation of the tool [79]. In addition to the morphological data, the dataset was also enriched by classifying the text paragraphs based on content. The paragraphs were tagged as lead paragraphs, text paragraphs, family relation paragraphs, or references to separate them from each other based on their location and content in a document. The data about the document structures, content, and text morphology was then utilized as an input for many other enrichments, such as NEL, and subject indexing.

The NEs (e.g., notable people, places, organizations) and keywords were extracted and added into document metadata with a custom property depending on the dataset. This is described in the publications I and VII. For automatic subject indexing and NEL, the AATOS tool was created. In the publication I, it was applied to the WarSampo dataset to extract the place and military unit names from text and added into the document metadata for each magazine article. The subject indexing was not done for the WarSampo dataset. The AATOS tool was also applied to the Semantic Finlex knowledge graph to perform only subject indexing and adding keywords linked to the original Finlex vocabulary (FinlexVoc) for the statutes of the knowledge graph. In BiographySampo (Publication VII), the keywords were generated using the same methods as in AATOS tool (TF-IDF and entity linking) in Semantic Finlex. However, the keywords were

[3] http://data.finlex.fi/

[4] http://www.finlex.fi/

linked to the General Finnish ontology (YSO)[5] to enable finding content using the same keywords from different databases that use YSO for subject indexing. The NE extraction for the WarSampo's dataset magazine articles succeeded in linking places and with the accuracy of 62.00% for places and 81.00% for military units (Publication I, Section 3.3). The subject indexing for Semantic Finlex was evaluated using R-precision and it succeeded with the accuracy of 45.45% (Publication I, Section 4.2). The subject indexing for biographies has not been evaluated as there is no gold standard to compare to.

The extraction of NEs for BiographySampo was achieved by building the NELLI tool that uses the knowledge graphs as source dataset that are created by the NLP pipeline. The NELLI tool is described in Publication V. In NEL, the NELLI tool utilizes numerous NER tools and the ARPA entity linking application. Similarly to [75, 27], the NELLI tool process has been broken into three tasks in annotating a text document corpus: NER, candidate searching, and NED. The tools FiNER[6] [201], LINFER, and ARPA [148, 120] are used with the BiographySampo dataset but due to configurability, other tools can also be added. The LINFER tool was developed to aid in the NER process. It utilizes the linguistic RDF data, i.e., information and Dependency Grammar relations. The disambiguation strategy utilizes a voting scheme [186, 58] variant where each tool has a vote based on its interpretation about the same piece of text. Here, the scheme also took into account the number of times an entity was recognized by different NER tools as a certain type (e.g., a person name or a place name), the string length, and linkage to a domain specific ontology. In addition, the NER tools were selected to provide three different approaches to improve results of the applied voting scheme in NED. The NEL process was done by linking entities to BiographySampo's own internal collection of person and place names from the dataset and the links were provided in the results for entities. Based on the best scores the application returns the results in RDF format. The automatic annotation pipeline managed to extract and link 74.00% of the person names and 84.00% of the place names correctly as shown in Publication V, p. 10-11.

In addition to previously presented enriching tools, the HENKO ontology for person names and tools using it (e.g., Person Name Finder and Gender Identification Service) were created. The ontology and the tools are described in Publication VI. The HENKO dataset is collected from multiple sources: The Finnish Digital Agency[7] (FDA), Norssi High School Alumni on the Semantic Web [95], BiographySampo [96], and AcademySampo [138][8]. From these datasets, the largest one is from the FDA that

[5] https://finto.fi/yso/en/

[6] https://github.com/Traubert/FiNer-rules/

[7] https://dvv.fi/en/individuals/

[8] https://seco.cs.aalto.fi/projects/yo-matrikkelit/en/

publishes modern Finnish name data as open data in the governmental publication portal avoindata.fi[9]. AcademySampo contains names of university students from 1640 to 1899, the Norssi Alumni dataset records students from a Finnish Normal Lyceum from 1867 to 1992, and the BiographySampo data contains names of notable Finns from the 3rd century to present time. The dataset was also enriched by utilizing NLP methods, such as LAS tool's lemmatization, hyphenation (see Publication VI, Section 2), to identify matronymics and patronymics, extraction of suffixes and nobiliary particles, and finally linking names to knowledge graphs. The enrichments were evaluated (Publication VI, Section 5) and added into the HENKO knowledge graph. The accuracy of identification of matronymics was 87.27% and for patronymics 94.42% for sample size of 1000 randomly selected names. The F1-score for extraction of suffixes 92.78% and particles 100% was calculated also for 1000 randomly selected names. The linking of names to various knowledge graphs varied depending on the target graph; roughly 23600 names were linked to Wikidata, 2500 to DBpedia, 785 to YSO places ontology[10], and 30 to AMMO ontology [64].

The tools, Person Name Finder and Gender Identification Service, utilize HENKO (model depicted in Publication VI, Fig. 1) and help to enrich and add information to a dataset. HENKO contains statistical information from the Finnish Digital Agency's records of names and combines it with name usage statistics from the different Sampos. It can be then used by Person Name Finder and Gender Identification Service to extract person names and identify gender for the name in addition to linking the names to HENKO. This can be then used to enrich datasets and to provide recommendations for the user. A common problem in NED is that the person names are mixed with, for example, place names [188] because many place names can be used as family names. The same applies also to other names, such as vocation names when written with a capital letter, e.g., in beginning of a sentence. For this reason the names in HENKO were linked initially to ontologies with place and vocation names. Thus, the linked names can be used to identify names that can also be places or vocations, e.g., the family name *Pappi* (in English priest). The accuracy of Person Name Finder has not been yet evaluated whereas the Gender Identification Service's algorithm has been evaluated to function in 97.70% accuracy (Publication VII, Section 5).

Lastly, a method for disambiguation of place names was developed and tested with old Finnish newspapers. The algorithm and results are described with more details in Publication IV. The NEL for OCR'd newspaper articles used a similar modular architecture as in AATOS and NELLI but it utilized Stanford NER [59] and ARPA [148, 120] tools. As shown in Publication IV, Section 3, it managed to link 355 correctly out of 672 place

[9] `https://www.avoindata.fi/en/`

[10] `https://finto.fi/yso-paikat/en/`

names. The result of the Stanford NER contained numerous false positives and OCR errors. In addition, 21 errors were caused by the disambiguation strategy used with the ARPA tool. Also, the process had issues with linking due to inflecting and baseforming Finnish words and place names that did not exist in the ontologies for place names (e.g., smaller Swedish towns, villages) but both cases were rare.

3.3 Explorative Applications and Data Analysis

After transforming and enriching the data, it can be used to build applications, for example, for data analysis. In this section, the results and applications related to research question 3 are presented in the first section. The second section describes the use of the applications for data analysis and presents results related to research question 4.

3.3.1 Tools for Search, Exploration, and Data Analysis

Knowledge graphs of text document collections can be utilized to build various search, exploration, and data analytical applications. This section answers research question 3: Can the semantic data enriched with linked named entities, keywords, and linguistic data be utilized to build search, exploration, and data analytic tools and applications for prosopographical research? The main research objectives of this section are OBJ 4 (Applications for search, exploration, and data analytics) and 5 (Promote generalizability of tools and applications to other CH collections). The publications I, III, and V explore the building of search, exploration, and data analytics pilot systems based on the data.

Based on the transformed and automatically generated data, Sampo systems can be supplied with tools for search, exploration, and data analytics of Finnish texts. The search applications, such as the faceted search application in the Sampo model, can be supplied with new facets. In BiographySampo and WarSampo portals the users can search for biographies using the keywords that have been added into the data. The WarSampo portal has also facets for linked mentions of military units and places that can be used to browse articles that mention them (Publication I, Fig. 2).

Based on the enriched data, the Sampo systems can be also supplied with network and reference analysis tools that can be used to explore and study the content of the datasets to get more insight about the people mentioned in and their demographics, similarly as has been done in BiographySampo. The user can see word frequency listings and generate networks for individuals using various tools and libraries. By utilizing networks analysis, the references of people extracted and linked with NLP-pipeline and Nelli can be used to visualize their connections from egocentric (Publication V, Fig. 1) and sociocentric points of view (Publication III, Fig. 6; Publication V, Fig. 2). These networks have been created using network analysis libraries and tools, such as Gephi and NetwrokX[11], to generate a reference network which is similar to citation networks [212]. In BiographySampo, the network view also has numerous toggles to control the construction of the network based on data (e.g., HTML links, automatically extracted person references, or both) and network visualization methods to decide the node size (network metrics) or color (person's gender, vocation) in the network. Also, a reference analysis application was developed to study the connections in their context given a list of sentences where the mentions occur (Publication V, p. 7-9). In BiographySampo, the application shows the sentences mentioning the biographee and how other people are mentioned in the biographee's biography. With similar logic, two bar charts were built (Publication V, Fig. 3-4) to show how many times the biographee is mentioned in history in other biographies and vice versa using birth years of the biographees by decade. The plots demonstrate the relevance of a biographee in Finnish history through the references.

In addition to the networks and reference analysis application, the references to NEs (e.g., people and places) are also shown in the contextual reader demo that can be created based on the knowledge graph and used to show NEs in text documents (Publication V, Fig. 5). There the user can view the mentions of people and places as links that are highlighted in the text. This is similar to the original national biography portal but it shows all links instead of just the first mentions and also mentions of places. The contextual reader is added to enhance the reading experience [163, 140] by providing information about the linked entities in both BiographySampo and WarSampo systems.

The linguistic data can be utilized to build applications to study word usage in text documents. The BiographySampo portal contains an application that can be used to view individual text document's most common words for different POS categories (Publication III, p.8-10). In addition to studying individual texts, the data is used to build applications to study the texts of different prosopographical groups. The data is used to create multiple statistics based on the linguistic information. The user can view the most common words for a group in addition to comparing word listings of different prosopographical groups, such as male and female politicians. It also contains total word counts of biographies on a timeline and for different vocational groups. These tools enable, for example, to study the language usage in the biographees of musicians or writers and compare them.

[11] `https://networkx.org/`

3.3.2 Knowledge Discovery for Digital Humanities

Knowledge graphs of text document collections can be used for data analysis. This section answers to research question 4: Can semantic data and its applications enable new data analysis methods in biographical and prosopographical research? The main research objectives of this section are OBJ 6 (Facilitate linked-data-based biographical and prosopographical research) and 7 (Evaluate the applicability of the data in proof-of-concept systems). The Publication VII explores the data analytical features of the BiographySampo dataset and describes how a linked data infrastructure can be utilized to study a dataset. The transformation of data into RDF and enriching it by linking enables building of search and exploration systems that help the users to view data from different points of view and to study it. The applications of BiographySampo that utilize the linguistic data can be used to study the language usage of the authors and the biographies of different prosopographical groups. Similarly, the network analysis, reference analysis, and the contextual reader application can be used to study the individuals or demographic groups in the dataset. They also complement each other and provide different levels of information to the reader, thus facilitating novel and diverse usage of semantically enriched dataset in biographical and prosopographical research.

In addition to the applications created for BiographySampo, the data was also utilized to create further data analytical visualizations to analyze and to learn about the collection. The data analytical visualizations were created by utilizing the cleaned and structured data in a SPARQL endpoint through a Google Colaboratory environment. The analyses were created using the data about biographies and their authors. The biographical data analytics were focused on the biographees as individuals and groups. For example, the analytics about the biographies indicated that most people in the texts were men from the fields of politics, culture, economics, and science. Approximately 10% of the biographies were about women of whom most worked in the same fields as men (Publication VII, Table 1). The analytics revealed also that the female biographees were often described with lesser personal details about their vocations (e.g., they have only academic degree title). The accuracy and rigor of the source dataset impacted the analytics. It can be hard to model and visualize data that contains inaccurate data, e.g., vague birth and death years of the biographees (Publication VII, Fig. 4) or vocational information (Publication VII, Fig. 8).

The applications of BiographySampo also visualized phenomena that required closer examination using Colaboratory. As an example, the inspections of the biographies using the linguistic analysis tool highlighted that women had more family related terminology in their biographies. The closer examination emphasized this and that men had more terminology related to economics, war, and religion in their biographies (Publication VII, Section 3.6). The analysis of HTML links visualized in the network view of BiographySampo also revealed that male biographees referenced more other males while females had a more evenly distributed set of references to both genders (also calculated in Publication VII, Section 3.4.2). This highlighted the choices made by the authors based on their information sources and guidelines [219, 218]. The linked data infrastructures and the data analytics created based on the dataset help scholars to understand the document collection and to identify its biases. However, it requires understanding of the source dataset, data transformation, and enriching processes. This kind of data literacy helps to evaluate the dataset's usefulness in biographical and prosopographical research.

3.4 Summary

In this section the results are summarized by revisiting the research questions presented in Section 1.4.

1. **Can a data model using existing models (CIDOC, NIF, DC-Terms) model text document structure of the cultural heritage domain?** A data model has been created using existing metadata schemas that are supplemented with a custom vocabulary for describing text documents, their structures, and content. By describing document structures in the model, the data can be used to study and enrich different parts of the documents, e.g., text paragraphs and their content. The addition of relations between different structures enables querying text data for different NLP tasks and using data analytical tools.

2. **The production of semantic data from text**

   (a) **Can a pipeline be built to transform Finnish documents and their collections be transformed into semantically rich and machine readable format?** In order to transform Finnish document collections into RDF, an NLP pipeline was created to transform documents of a collection in accordance with the data model. The pipeline preserves information from the documents and extracts document structures and existing markup from the text and transforms it into RDF.

   (b) **Can NLP-based methods can be used to build generilizable tools to enrich knowledge graphs for different document collections?** Knowledge graphs can be enriched with content describing metadata to facilitate knowledge discovery. For this purpose, tools were created to provide knowledge graphs with linked NEs and keywords. The NLP methods are used in extracting and linking people, places, and other entities to different ontologies. The document metadata and text provide context for the extraction and linking tasks. By creating configurable tools that can be used in different contexts, satisfactory results can be achieved for the tools.

3. **Can the semantic data enriched with linked named entities, keywords, and linguistic data be utilized to build search, exploration, and data analytic tools and applications for prosopographical research?** Search, exploration, and data analytical systems can utilize data extracted and produced based on natural language texts. The enriched data can be used to study and compare different prosopographical groups in the dataset. In addition, the search and exploration options that have been added through data enrichment enable browsing datasets by utilizing keywords or finding mentions to a place or a person in multiple different text documents. These applications can enable knowledge discovery.

4. **Can semantic data and its applications enable new data analysis methods in biographical and prosopographical research?** Using the transformed and enriched data, we can build data analysis tools. The tools can be used to study document collections and their provenance. In addition, the linked data services can be used to query results for different prosopographical questions. The biases of the document collections can be identified by using network analysis, reference analysis, and linguistic analysis. However, the interpretation of the results requires data literacy and understanding how the dataset has been produced before it can be utilized for biographical and prosopographical research.

4. Discussion

Traditionally in computer science, comparative evaluations of software artifacts are made to evaluate their effectiveness and usability. For example, tools and applications using NLP and IR techniques, have methods and algorithms for evaluating them. However, software artifacts, such as models, pipelines, and methods, presented in this thesis can be difficult to evaluate traditionally. The evaluation of systems in Semantic Web research area [15] is hard because their usefulness and usability depends on several factors, such as the quality of the used heterogeneous source data, the used software for data handling, and the user interfaces (UI) built for the data [237]. In other fields of computer science, like NLP or IR, there are often software artifacts and algorithms that can be used as basic building blocks when creating novel methods and applications. In Semantic Web research, there are in many cases few building blocks available and often everything has to be designed for every application. This adds complexity to the design and development processes of Semantic Web-based applications because a level of maturity and high-quality is required from them before they can be applied.

To evaluate the research of this thesis, the following criteria, that have been proposed by Burstein and Gregor [28] have been applied: 1) theoretical significance and practical significance, 2) internal validity, 3) external validity, 4) objectivity, and 5) reliability. In addition, the proof-of-concept systems have been utilized extensively to prove the usefulness and usability of models, tools, methods, and applications created in this thesis based on real world data to comply with the principles of action research. Mature evaluation methods typical for NLP applications have also been utilized when possible. In this chapter, the research of this thesis is evaluated with the presented criteria. Lastly, future research recommendations are presented.

4.1 Theoretical Implications

The theoretical implications of the thesis relate to the topics presented in Chapter 2. Here, the contributions of each research question are reflected against the research literature of these topics.

4.1.1 Modeling Document Collections

Can a data model using existing models (CIDOC, NIF, DC-Terms) model text document structure of the cultural heritage domain? (RQ 1)

Modeling of document collections has been done before in different projects. In this thesis, the document corpus collection has been modeled in RDF using mainly the NIF model, as described in Section 3.1. The main contribution of this modeling work is to create a model for representing a Finnish document collection in RDF format using existing ontologies and supplement them with new properties and classes for describing document structures and content, such as links in HTML documents. Prior to this work documents have been transformed to, for example, to TEI [169] and NAF [243, 60] formats that can also contain linguistic information and automatically extracted NEs. The NIF ontology provides a simple model for describing text structures. The NIF ontology was supplemented with other ontologies, such as CIDOC CRM, DC Terms, and a custom vocabulary.

The custom vocabulary was supplemented with support for describing text categories, extracted HTML markup, and automatically generated annotations, such as NEs and keywords. The NIF ontology has some support for annotations [26, 81] but to preserve the HTML markup in the text collection, new and descriptive classes were created for them. Also, since the BiographySampo dataset used in the study contained different kinds of annotations, it was important to separate them to improve their usability later in data visualizations and analytics. Similarly, the NEs were modeled differently than in NAF due to adding provenance information about them (confidence score and what features attributed to it, e.g., NER methods and their linkage).

4.1.2 Production of Semantic Data from Text

Can a pipeline be built to transform Finnish documents and their collections be transformed into semantically rich and machine readable format? (RQ 2a)

In modeling, the aim was to model the text document collection dataset to describe the documents and their collection metadata as it is presented in the original source material. The main goal of the NLP pipeline is to transform the document collections into a semantically rich format in accordance with the presented model. In this thesis, the transformation of the document collection has been described in Section 3.2.1. Prior this work, there has existed pipelines to transform texts to knowledge graphs, e.g., in FRED [192], NewsReader [200], and BiographyNet [62]. The pipeline created in this thesis is a modular pipeline made for Finnish texts. What separates it from other pipelines is that the whole document in text, HTML, or RDF format, word by word, is transformed into fine-grained RDF form, recording also full linguistic information of the texts and their structures.

The pipeline uses NLP methods to identify, extract, and transform information, such as text structures and formatting. It is similar to pre-existing NLP pipelines and toolkits offered to DH to ease their work [121, 182, 136, 6]. The contribution here is that the NLP pipeline also transforms pre-existing annotations and identifies different layers of text structures that are present in the text but are not always explicitly expressed. For example, it picks HTML annotations embedded in the text, such as HTML links and formatting, and transforms them into the RDF representation. The RDF representation is supplemented with structural information, such as location of the annotation in text. The extracted information makes the dataset more usable; the text is cleaned from HTML markup and the markup is added into the data for later usage, such as in data analytical applications.

In addition to these contributions, a Reasoner module was developed to supplement the CoNLL-RDF tool [33]. It uses the results of the CoNLL-RDF tool and infers missing information, such as connections between text structures that are part of the NIF model, and adds them to the dataset. This contribution completes the pipeline and improves the data usability.

Can NLP-based methods can be used to build generilizable tools to enrich knowledge graphs for different document collections? (RQ 2b)

The NLP pipeline produces linked data that can serve as a starting point for many enrichment processes. The main goal of the enrichment is to supplement the knowledge graph with additional information about the document collection. In this thesis, the main contribution here is the tools and resources created for the enrichment process. The enriching tools and resources created in this thesis have been described in Section 3.2.

One of the contributions in subject indexing is the AATOS subject indexing and entity linking tool for Finnish texts. Similarly to DBpedia Spotlight [163], it can be given text and configured to pick up NEs that exist in a given ontology but also keywords like in the Annif tool [221] that was created later. There exists a number of tools for subject indexing Finnish texts utilizing several approaches from machine learning to other NLP algorithms. AATOS is a highly configurable tool that is based on linked data and the simple NLP method TF-IDF. Entity disambiguation is based on configurations of the tool, such as ordering of the ontologies that are used as targets in linking to prioritize entities based on type, e.g., prioritizing people over places, because people can have family names that are also place names. Unlike Annif, the AATOS tool can also identify NEs from texts.

In addition to transforming the BiographySampo's documents to RDF, the words were in parallel enriched with linguistic information like in NewsReader [200] and BiographyNet [62] datasets. The purpose in this thesis was to use the linguistic information in a variety of tasks directly by querying it from a SPARQL endpoint. In the case of BiographySampo, the linguistic and morphological information in the dataset was used in NLP tasks instead of repeating the morphological analysis phase for each task. For NEL, NELLI tool was developed to enrich the dataset with linked NEs. The contribution of NELLI is that it utilizes a disambiguation scheme based on a voting scheme [186, 58]. In addition to using voting of different NER tools, NELLI takes into account the entity length, linkage, and NE type by enabling earning points for candidates that overlap but are, for example, longer or that are linked to an user-specified ontology. Another contribution is the LINFER tool that was implemented to supplement the entity extraction in NELLI by utilizing the results of morphological analysis and the Dependency Grammar relations as context. The NELLI tool is able to extract and link NEs with the help of FiNER, LINFER, and ARPA tools with satisfactory results in entity extraction. The results of NER in contrast to the new FinBERT [145, 241] tool were poorer. However, the system enables connecting other tools to it and in future it can be possible to add new tools to aid in entity extraction and linking.

In addition to the previously mentioned NEL and subject indexing tools, the HENKO ontology resource was created. It is utilized in Person Name Finder and Gender Identification services. The HENKO ontology is a unique resource that shares some similarities with Wikidata in regards to modeling names. However, unlike Wikidata, the ontology contains statistical information about name usage and provenance information about the data sources where the information is retrieved. The contribution of the services utilizing HENKO is that they use the unique information in HENKO for providing services for Finnish language environments. The Person Name Finder is a NEL tool for identification of person names. Its contribution is in linking person names to HENKO and providing information about the possible ambiguity by informing the user if a name in HENKO is also linked to a place or vocation name.

Lastly, an algorithm for disambiguating and linking geographical names was developed. Unlike prior work in geographical named entity disambiguation [71, 89, 65, 242, 29], the algorithm utilizes the metadata as context for linking names of smaller places into an ontology, utilizing coordinates to calculate the proximity of the location to a local magazine's publication place. The results of the methods were impacted by the OCR quality of the texts but it seemed to work for most places. The errors were mostly related to OCR errors. However, a formal evaluation is still needed to compare it to other similar works.

4.1.3 Explorative Applications and Data Analysis

Can the semantic data enriched with linked named entities, keywords, and linguistic data be utilized to build search, exploration, and data analytic tools and applications for prosopographical research? (RQ 3)

NLP methods have been used to transform and enrich existing datasets with novel information that can be used for a variety of applications. Here the enriched data is used to contribute to three types of applications; search, exploration, and data analytics. The applications are described in Section 3.3.1. The first contribution uses the enriched data in faceted search type applications where users can search documents that mention extracted people or place mentions or keywords. Automatically extracted data has been used in systems, such as ePistolarium [195] and Europeana [67, 183], to build search applications to improve finding of information from vast document collections. In Finland similar applications have surfaced to the field only recently as traditionally systems have relied on manually annotated documents. In this thesis, extracted mentions and keywords are linked to a given controlled vocabulary.

The second contribution of this thesis is related to applications utilizing the knowledge graph for data exploration and visualization. Here, HTML links and automatically extracted NEs from the text are used in the contextual reader, network analytical, and reference analysis applications to help the user to explore the dataset. Similarly to Wikipedia, the contextual reader application utilizes the linked mentions in the text as highlighted links. In contrast to WarSampo's Kansa Taisteli magazine article view's contextual reader [151], the application in BiographySampo shows preprocessed annotations and puts more effort on disambiguation than the CORE application in WarSampo. Similarly to [247, 179, 133, 25], the network analysis applications can be used to browse connections between people but here using the toggles based on data and network visualization methods. The reference analysis is a novel contribution on the networks, by giving meaning, explanations, and context to the links in the networks. It is similar to the contextual reader but instead of the whole text, it shows NEs in mentioning sentences. The visualization of NEs is similar to the visualization of words in the KORP tool [21] but instead of mapping mentions of NEs to a map, they are placed on a timeline plot.

The third contribution related to data analytical applications is the linguistic analysis tool available in BiographySampo. It is a novel tool for studying biographies and texts that gives a glimpse into their content to get more insight about individuals and the demographic in the dataset. It is similar to the VoyantTools [209] online application but the tool is integrated with BiographySampo. Similarly to VoyantTools, it shows word count listings of the most common words for individual biographies and groups of biographies in separate views. By supplementing the basic information about the biographees with text document structures and linguistic information, this tool can be used to study and compare prosopographical groups within the dataset. In addition, the dataset can be used to present linguistic statistics on word usage in texts which gives more context to the word listings.

Can semantic data and its applications enable new data analysis methods in biographical and prosopographical research? (RQ 4)
Biographical text document collections can be used as an artifact to study the underlying world from a historiographical perspective. They reflect their own time, the editorial values and biases in selecting the biographees, and the authors' standpoint from a linguistic point of view. The data analytical applications, created as part of BiographySampo, serve as proof-of-concept systems that give a glimpse into the underlying data and provide novel tools for scholars to use when evaluating the usability of the dataset for their research purposes. The use of the data to analyze the NBF dataset is described in Section 3.3.2.

In addition to the applications in BiographySampo, the underlying cleaned dataset provides DH scholars a starting point for further data analytics [150, 16]. Biographical collections have been analyzed and studied before in Finland [132] and around the world [246, 62, 134, 16]. In addition, there is analogous research that uses Wikipedia articles as the source dataset [111, 164]. The approach presented in this thesis contributes by giving scholars biographical and prosopographical tools for studying individuals and groups of people. The tools merge quantitative approach and distant reading techniques [109] with the qualitative approach, often based on close reading, that is common to biographical research.

The analytical applications and visualizations identify the characteristics of the dataset through a variety of statistics. The features and biases, such as the low number of women in the data, are similar to what can be found in other European national biographical dictionaries [246, 16]. However, the analytical visualizations presented in this thesis extend these analytics by describing the dataset and also takes into consideration how the data has been produced [150, 132]. The use of linked data infrastructure created for BiographySampo enables knowledge discovery and various data analytical visualizations. In addition to learning about the demographic through the statistics, the user can also study connections be-

Table 4.1. The objectives and the corresponding solutions presented in this thesis.

Objective		Solutions
OBJ 1	Model for text document collections	A general data model for describing text document collections' documents and their inner structures. To harmonize the document collection with other collections, the model uses mainly existing metadata schemas that are extended with a custom vocabulary. The model can be used to build data analytical applications and visualizations.
OBJ 2	Pipeline for transforming text to linked data	An NLP pipeline for transforming text document collections into linked data in accordance with the data model. The NLP pipeline preserves information from the original text document collection to maintain usability and integrity of the data.
OBJ 3	Facilitate knowledge discovery	Knowledge discovery is supported with tools that can be applied for enriching linked-data-based document collection knowledge graph. The tools can be used for NEL and subject indexing. The NLP pipeline also enriches the text document collections with results of morphological analysis as RDF.
OBJ 4	Applications for search, exploration, and data analytics	Pilot applications were created for search, exploration, and data analytics. They can be used for analyzing datasets and their content. The distant and close reading methods are also utilized in the data analytical and exploratory applications.
OBJ 5	Promote generalizability of tools and applications to other CH collections	The tools and applications created in this thesis are generic and can be used mainly on knowledge graphs that use the data model presented in this thesis. The tools and the NLP pipeline support various input forms, are highly configurable to the environment, and use a modular architecture.
OBJ 6	Facilitate linked-data-based biographical and prosopographical research	The knowledge graphs based on the created data model provide infrastructure for biographical and prosopographical research directly through a SPARQL endpoint, or by using the created pilot applications for search, exploration, and data analytics.
OBJ 7	Evaluate the applicability of the data in proof-of-concept systems	The feasibility of the knowledge extraction kit to document collections has been demonstrated by the pilot applications and data analytical visualizations created based on the knowledge graphs. The datasets and applications have been evaluated for each application where appropriate.

tween single biographees through the network visualizations and reference analysis tools. The actor knowledge graph is published partially openly (for deceased individuals) and can be queried with SPARQL to study the demographics. Unlike in [246], the LOD service enables further querying, analyzing, and visualizing the data.

4.1.4 Summary

The theoretical implications of the resources, methods, and tools that have been presented in this thesis are summarized in Table 4.1, by re-visiting the objectives for the knowledge extraction toolkit defined in Section 1.3.

4.2 Practical Implications

The practical implications of the contributions of this thesis are presented in this section. The results and contributions of this thesis are presented Chapter 3. Following subsections elaborate the practical benefits of the solutions and their implications in more detail.

4.2.1 Modeling Document Collections

Can a data model using existing models (CIDOC, NIF, DC-Terms) model text document structure of the cultural heritage domain? (RQ 1)

The successful modeling of document collections improves the usability of the knowledge graph utilizing it. In Section 3.1 the modeling of document collections is described in more detail. The data model is generic and can also be used for other biographical or historical document collections in addition to BiographySampo. It enables not only interlinking of collections but also search, exploration, and data analytical applications similar to applications made for BiographySampo. Therefore, the data model provides a starting point for scholars, computer scientists, and others who utilize it in their document collections to build data analytical applications.

The modeling of document structures enables the use of the data as a starting point for various NLP tasks. In this thesis, the NIF model used in modeling document structures was expanded with morphological data and with extracted HTML annotations from source documents. The data model was also expanded with information about the text paragraph types to identify the differences between bodies for text, such as text and semi-structured paragraphs. This makes the model usable for text analysis and NLP tasks, because, for example, the text paragraphs can be separated and the lemmatized form can be queried with a SPARQL query. The results can be used as an input for NLP tasks to enrich the dataset with, for example, keywords, NEs, or to create visualizations, such as tag clouds.

The NIF model was also extended with annotations, such as NEs and extracted HTML links. For NEs, this enables querying them and their textual context, e.g., querying documents mentioning the person or words in close proximity to the mentions of a person or place. The model for NE contains provenance about how and from which source vocabularies it was selected. This improves transparency of the data by providing simple provenance information to its users. The HTML markup was removed from the text and moved to the data to provide users with cleaned text documents. The HTML links to other biographies were given their own class to distinguish them from NEs and other extracted HTML markup. Similarly to NEs, they can be used to find passages of text surrounding the HTML links between two documents. The model provides users with the possibility to study the texts using manually made HTML links and linked NEs.

4.2.2 Production of Semantic Data from Text

Can a pipeline be built to transform Finnish documents and their collections be transformed into semantically rich and machine readable format? (RQ 2a)

The NLP pipeline transforms text document collections into RDF. In Section 3.2.1 the transformation process is described in more detail. The pipeline's goal is to preserve information from the text collections to create linked data infrastructure for the document collection. The results of the pipeline provide users with a knowledge graph that can be used to query cleaned text or parts of it directly, such as paragraphs or sentences, to use it for NLP tasks. NLP methods can be used to enrich the data, for example, with linguistic information or NEs. The pipeline's results can also be used to build data analytical applications, similar to applications in BiographySampo. The documents can be studied by creating data visualizations based on the extracted HTML links or to provide statistics about document word counts.

The NLP pipeline is generic and it can be used directly for other Finnish document collections in text, RDF, or HTML format. The pipeline is built from different components for analyzing and transforming Finnish language texts to RDF. It utilizes pre-existing tools and methods for linguistic analysis and RDF conversion to build a knowledge graph for text document collections. The pipeline architecture is modular and by replacing software modules, it can be used, for example, tools for other languages that produce similar data.

Can NLP-based methods can be used to build generilizable tools to enrich knowledge graphs for different document collections? (RQ 2b)

Text document collection knowledge bases can be enriched with NLP methods. The Section 3.2.2 describes the enriching of the knowledge graphs. In this thesis, the document collection knowledge graphs have been enriched with NEs, keywords, morphological information, and text categories. For this purpose the tools AATOS, NELLI, Person Name Finder, and LINFER and custom scripts have been created. In addition, the Person Name Finder relies on Gender Identification Service and HENKO ontology that are used to extract contextual information about the person names.

During the transformation of text to RDF, the NLP pipeline enriches the document collections with results of morphological analysis. The data can be queried for linguistic information and lemmas directly from the knowledge graph, for example, to perform various NLP tasks similarly to

NELLI. The NELLI tool can be used by scholars and researchers for NEL to enrich the existing metadata. It produces the results in RDF and links the entities to the underlying document structures. This improves access to the documents, their sentences, paragraphs, and words that the entities consist of or are part of. NELLI can be configured to link the entities to different ontologies and shared vocabularies to improve interlinking of the documents internally and externally to other document collections. The NELLI tool demo has been made available for the public[1] [223].

Similarly to the NELLI tool, AATOS can be utilized for different Finnish text document collections to enhance their metadata with keywords and NEs. AATOS takes as input text document files and adds its results to the preexisting RDF-based document collections. It can be used, for example, by librarians, computer scientists, and curators. The tool is available publicly on GitHub[2].

The HENKO ontology resource utilizes the name data from various source datasets and the information is used to create a knowledge graph about person names. It is a unique resource for computer scientists and scholars who can use it to build applications, train machine-learning-based tools, and to study person name datasets. The ontology is available as LOD on the Linked Data Finland (LDF.fi) platform [100]. It provides a public SPARQL endpoint[3] and a dataset description page[4] with general documentation. The human-readable data model documentation is available for users[5]. In addition to Linked Data Finland platform, the ontology is also published in the ONKI Light service[6], where it can be searched and browsed using SKOSMOS[7], a web-based SKOS browser. The HENKO ontology has been used to build applications, such as Person Name Finder[8] and Gender Identification Service[9], that have been used to extract information and enrich datasets. Similarly to HENKO, these tools can be used by, for example, computer scientists, scholars, and librarians to enrich their datasets. These applications will be included as part of the SeCo Text Annotation Service along with Application programming interface (API) descriptions of the services in future.

[1] http://nlp.ldf.fi/appi/

[2] https://github.com/SemanticComputing/AATOS/

[3] http://ldf.fi/henko/sparql

[4] http://ldf.fi/dataset/henko/

[5] http://ldf.fi/schema/henko/

[6] http://light.onki.fi/henko/en/

[7] http://skosmos.org/

[8] http://nlp.ldf.fi/name-finder/

[9] http://nlp.ldf.fi/gender-identification/

4.2.3 Explorative Applications and Data Analysis

Can the semantic data enriched with linked named entities, keywords, and linguistic data be utilized to build search, exploration, and data analytic tools and applications for prosopographical research? (RQ3)

The semantically rich linked data created from natural language texts can be used to build pilot applications. In Section 3.3.1 the search, exploration, and data analytical systems based on linked data created with NLP methods are described. It shows how the enriched linked data can be used to build pilot tools and applications to study the document collection and its phenomena. RDF transformation enables easy access to the data through endpoints by utilizing SPARQL query language. The transformation of text to RDF created a large amount of data and therefore sometimes even simple queries can cause issues when using it. Currently, it can sometimes take long time for the database to process the queries and produce results. In the future, the processing would benefit from even more helper properties and relations in addition to reconsidering hosting of the dataset in a database that supports big data. At the moment, in order to use the data, the user needs to under stand how to operate with the database to get most out of it. Thus, the data can be utilized in applications for prosopographical research by different user groups, such as DH scholars or computer scientists. In this case IR and data visualization applications were created on top of the knowledge graph. This enables using the enriched metadata in a faceted search application to study how a person or certain people are written about or the use of vocabularies for a specific group of people (e.g., by vocation, gender, or birth place) in biographies.

The extracted and enriched document metadata can be utilized in faceted search applications also for finding documents related to, for example, topics, themes, and extracted NEs. They also connect the documents to other document databases and help to find related documents, e.g., documents about a topic, place, or a person. This feature can bridge together different databases and ease IR. For these purposes applications can be made to study the references to NEs, e.g., sociocentric or egocentric network visualization, reference analysis tools to study the context where mentions are made, and contextual reader tools to visualize the NEs in texts. These tools can be useful for scholars to identify and interpret how people are talked about and their importance. They can also be a useful part of education to help to understand the topic and its context by providing more information for the readers about mentioned NEs. In addition to the NEs and keywords, the linguistic information can be utilized to analyze the language usage in the texts. Applications similar to the Voyant Tools [209] can be built on top of the data to analyze the text content and to visualize it, for example, with tag clouds. The applications based on linguistic data can be useful for computational sociolinguists as they can give quickly a glimpse into the data, e.g., how different prosopographical groups are described in the collection.

In 2018, the NBF's biographical collections were re-published in the BiographySampo portal [97]. The portal includes the pilot systems for network analysis, reference analysis, contextual reader, and linguistic analysis.

Can semantic data and its applications enable new data analysis methods in biographical and prosopographical research? (RQ4)
The semantically rich knowledge graphs can be used for data analysis to learn about the features of the knowledge graphs. This is described in detail in Section 3.3.2. The proof-of-concept data analysis applications based on extracted knowledge from the document collections highlight existing relations between themes, topics, and agents described in the documents. The distant and close reading applications, such as network analysis, reference analysis, linguistic analysis, and contextual reader applications, enable inspecting the topic from different perspectives. Therefore, in addition to gaining understanding of the networks and their relations based on the link analysis, the users can also discover new themes and connections about the target. The contextual reader applications show the entities in a broader perspective with other entities. In addition to highlighting relations between documents, these applications, and especially the linguistic analysis application, highlight the choices made by the authors and editors of these collections. For the scholars who use this data for research, this information is valuable to evaluate the strengths and weaknesses of the dataset.

In addition to the data analytical tools, the data can be used through the SPARQL endpoint. The data can be used, for example, not only to repeat and validate existing visualizations but also to create more complex data visualizations and analytics about the dataset. The data can be used by scholars to study phenomena present in the data via visualizations of the data analytical tools. It is available to scholars, computer scientists, and others through a SPARQL endpoint and can be utilized flexibly by using tools, such as Yasgui for SPARQL, or Jupyter and Google Colaboratory by Python scripting to do linguistic or network analysis.

Thus far, the BiographySampo portal [97] has had approximately 40 000 of end-users on the Web since its publication in 2018. The applications based on data extracted from the biographical text collections are available for scholars and the public. Similarly, the knowledge graph is partially available through the endpoint[10]. The document texts and their linguistic information is not publicly available and their usage requires negotiations with the copyright holders.

[10] `https://ldf.fi/nbf/sparql`

Table 4.2. Contributions and implications of the knowledge extraction toolkit.

Contribution	Implications
Data model for text document collections	A generic data model provides an easier starting point not only for various NLP tasks but also for scholars, computer scientists and others. It can be used to transform document collections to RDF to build linked data infrastructure. The linked data infrastructure can be used to build Semantic Web-based applications for end users.
NLP pipeline	A modular NLP pipeline that can be used to transform the document collections in accordance to the data model. It extracts and transforms document structures into RDF representation, enabling the use of clean data for visualizations and applications.
Tools for enriching knowledge graphs	Several generic and configurable tools were created for enriching knowledge graphs with linguistic information, NEs, and keywords. The linguistic data can be used as a starting point for NLP tasks, such as NEL and subject indexing. Also, HENKO was created as a resource for computer scientists and scholars who can use it to build applications, train machine-learning-based tools, and to study person names in datasets.
Applications for search, exploration, and data analysis	The linked data infrastructure was used to create pilot applications for search, exploration, and data analysis for prosopographical research. They can be used by scholars to inspect document collections using the pilot systems for close and distant reading.
Methods and infrastructure for analyzing knowledge graphs	The linked data services enable to repeat and validate existing visualizations in addition to creating complex data visualizations and analytics. The pilot applications highlight the choices made by the authors and editors of these collections. Together, the applications and the data can be used by scholars to study phenomena present in the visualizations of the data analytical tools.

4.2.4 Summary

The resources, methods, and tools created in this thesis are part of the knowledge extraction toolkit. The contributions of the toolkit and their practical implications are summarized in Table 4.2.

4.3 Reliability and Validity

The quality of research is evaluated by utilizing the concepts of reliability and validity. Reliability is about the quality control of the research process; it's about the consistency and stability over time, across researchers and methods. The overall research goals, objectives, and questions have been presented in Chapter 1. The research questions have been revisited in Chapter 3 after presenting the results of the thesis. The developed tools and prototype systems have been presented and covered in detail in Chapter 3 in addition to being presented in more detail in the related publications. The research objectivity is preserved by detailing the research methods, resources (e.g., used ontologies, tools, and their configurations), and describing the participating organizations. The research has been conducted following the principles of design science and the results have been reported regularly to the research community for evaluation. Lastly, the author of this thesis has no competing interests towards the presented research and does not recognize personal biases that could have impacted the research process.

The validity evaluates if the research process is well-founded and likely corresponds accurately to the real world. It is divided into internal and external validity [28, 30, 39]. The internal validity concerns the stated research objectives and the requirements of the developed systems, the alternative methods, and the limitations of the research. The developed prototype systems, models, and methods meet the objectives presented in Chapter 1. As proof of concept, they have been applied in real-world situations as has been showcased in Chapter 3 and in Sections 4.1 and 4.2 of this Chapter. The text collection of the BiographySampo was transformed into RDF and enriched using NLP methods. This was followed by building IR, and data analytical applications that could be used to study the text collection to understand its nature and to learn more about its subjects. The used models, created applications, and findings have been compared with relevant related work.

The research questions were formed based on research objectives and the developed artifacts meet these objectives and answer the research questions set in Chapter 1, as is discussed in Chapter 3. The enrichment of the knowledge graph was evaluated in the publications with established IR evaluation metrics, such as Precision, Recall, F-measure, and R-Precision [152]. To evaluate the data analytical tools, the analyses were done in collaboration with humanist scholars to understand better the phenomena behind data statistics and visualizations. The developed methods and tools provide data users and providers with insight [16] from what it is possible to learn more about the data and its production. The data analyses also contributed to the improvements of the underlying data model as many new features have been added (such as text type categorization) to enable easier data analytics.

The external validity refers to the generalizability of the research results and their congruence with preceding theory. The proof-of-concept systems and data analytics demonstrate the applicability of the developed methods, models, and tools in other similar systems. The use of generic vocabularies enable the use of the model in other text collections as well. These vocabularies and models have not been designed for a particular text collection but for different collections, and therefore the data model underlying the data analytics and systems is applicable to other collections as well. The tools built to enrich the knowledge graph (NELLI, Person Name Finder, Gender Identification Service, AATOS) are available online and can be also used to enrich other similar datasets that are based on linked data and Semantic Web technologies. Similarly, the NLP pipeline for transforming and enriching text documents has been compiled using existing tools that are available online to create text document collections as enriched RDF datasets that can be used in linguistic analysis or network analysis.

The data models, tools, and methods of this thesis have been designed by taking into consideration the previous research. The text document collection data model utilizes vocabularies created earlier by scholars. It takes inspiration from previous work on modeling text document collections and datasets. The tools presented in this thesis are built using architecture presented in earlier work and utilizing mature software modules. The methods used in this thesis are aligned with prior work and used in accordance with the theory and practise of previous works.

4.4 Future Research Recommendations

The research presented in this thesis opens up opportunities for improvement in several areas. The NLP pipeline has been applied to BiographySampo, but in the future, it could be applied to and tested with other Finnish document collections to enable data analytics. Similarly, the data model for document collections can be improved by including more metadata about the texts. For instance, the Finnish Turku dependency parser can be upgraded to a newer version and its results analyzed further to model compound words and other features of Finnish language. In addition, based on more fine-grained data, it would be useful to develop more corpus linguistic tools for DH scholars.

The enrichment applications could be also tested with other datasets and upgraded to use latest versions of software and new tools for NER and NEL. The Person Name Finder tool for data enrichment is still partially under construction. The tool includes links to systems containing people with similar names. These links could be utilized to let the user know of the person name best matching to the given name in other services to build a person linking tool. The name linking tool could be also utilized to extract data, such as titles or vocations, that can be included preceding of the name in text.

Extending the HENKO ontology to include information about compound words in names and name translations could also be useful. The compound-word-based names could be modeled into the dataset by utilizing models equipped to modeling language and its features in more detail, e.g., Ontolex-Lemon. The support for studying name changes and translations could also be improved by identifying and connecting names to their variants. In addition, the ontology can be enriched by linking names to other ontologies that can contain themes present in person names (e.g., birds, mammals, plants, natural phenomena). Similarly, the names could be linked to Finnish organization names to help users to identify that a name can be also an organization name or part of it (e.g., family name Pöyry and the company Pöyry Oy).

The services, tools, and applications of this thesis could be brought to their own portal from where they would be easier to access and test by scholars and software developers. Currently, such a portal is under construction and in addition to tools and applications it will contain their descriptions, API descriptions, and demo UI for testing the applications.

References

[1] BD2015 Biographical Data in a Digital World 2015. In *Proceedings of the First Conference on Biographical Data in a Digital World 2015, Amsterdam, The Netherlands, April 9, 2015* (2015), S. ter Braake, A. Fokkens, R. Sluijter, T. Declerck, and E. Wandl-Vogt, Eds., CEUR Workshop Proceedings.

[2] BD-2017 Biographical Data in a Digital World 2017. In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017, Linz, Austria, November 6-7, 2017.* (2017), A. Fokkens, S. ter Braake, R. Sluijter, P. Arthur, and E. Wandl-Vogt, Eds., CEUR Workshop Proceedings.

[3] ALKULA, R. From Plain Character Strings to Meaningful Words: Producing Better Full Text Databases for Inflectional and Compounding Languages with Morphological Analysis Software. *Information Retrieval 4*, 3-4 (2001), 195–208.

[4] ALLEN, J. *Natural Language Understanding*, 2nd ed. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA, 1995.

[5] ARDEN, B. W. *What Can be Automated?: The Computer Science and Engineering Research Study (COSERS)*. MIT Press, 1980.

[6] ARNOLD, T. A tidy data model for natural language processing using cleanNLP. *The R Journal 9*, 2 (2017), 248–267.

[7] ASHTON, J., AND KENT, C. New approaches to subject indexing at the British Library. *Cataloging & Classification Quarterly 55*, 7-8 (2017), 549–559.

[8] 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: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*. Springer, Berlin, Heidelberg, 2007, pp. 722–735.

[9] AYLETT, R. S., BENTAL, D. S., STEWART, R., FORTH, J., AND WIGGINS, G. Supporting Serendipitous Discovery. *Digital Futures (Third Annual Digital Economy Conference)* (2012).

[10] BACA, M., Ed. *Introduction to Metadata*, 3 ed. Getty Publications, Los Angeles, California, 2016.

[11] BASKERVILLE, R. L. Distinguishing action research from participative case studies. *Journal of systems and information technology 1*, 1 (1997), 24–43.

[12] BASKERVILLE, R. L., AND WOOD-HARPER, A. T. A critical perspective on action research as a method for information systems research. *Journal of Information Technology 11*, 3 (1996), 235–246.

[13] BERNERS-LEE, T. Linked data, W3C design issues. `http://www.w3.org/DesignIssues/LinkedData.html`, 2009. Accessed 3 April 2021.

[14] BERNERS-LEE, T., HENDLER, J., AND LASSILA, O. The semantic web. *Scientific American 284*, 5 (2001), 34–43.

[15] 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.

[16] BHREATHNACH, Ú., BURKE, C., FHINN, J. M., CLEIRCIN, G. O., AND RAGHALLAIGH, B. O. A quantative analysis of biographical data from Ainm, the Irish-language Biographical Database. In *Proceedings of the Third Conference on Biographical Data in a Digital World 2019, Varna, Bulgaria, September 5-6, 2019.* (2019), CEUR Workshop Proceedings.

[17] BIBER, D. *Corpus-Based and Corpus-driven Analyses of Language Variation and Use*. Oxford University Press, September 2012.

[18] 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.

[19] BLAXILL, L., AND BEELEN, K. A Feminized Language of Democracy? The Representation of Women at Westminster since 1945. *Twentieth Century British History 27*, 3 (2016), 412–449.

[20] BORGMAN, C. L. The Digital Future is Now: a Call to Action for the Humanities. *Digital Humanities Quarterly (DHQ) 3*, 4 (2009), 1–30.

[21] BORIN, L., FORSBERG, M., AND ROXENDAL, J. Korp - the corpus infrastructure of Språkbanken. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), May 21 - 27, 2012, Istanbul, Turkey* (2012), European Language Resources Association (ELRA), pp. 474–478.

[22] BÖRNER, K., SANYAL, S., AND VESPIGNANI, A. Network science. *Annual review of information science and technology 41*, 1 (2007), 537–607.

[23] BORST, W. N. *Construction of engineering ontologies for knowledge sharing and reuse*. PhD thesis, University of Twente, Netherlands, 1997.

[24] BROOKE, J., HAMMOND, A., AND HIRST, G. GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature, June 4, 2015, Denver, CO, USA* (2015), Association for Computational Linguistics, pp. 42–47.

[25] BROUWER, J., AND NIJBOER, H. Golden Agents. A web of linked biographical data for the Dutch Golden Age. In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017, Linz, Austria, November 6-7, 2017.* (2018), CEUR Workshop Proceedings, pp. 33–38.

[26] BRÜMMER, M., DOJCHINOVSKI, M., AND HELLMANN, S. Dbpedia abstracts: A large-scale, open, multilingual NLP training corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), May 23-28, 2016, Portorož, Slovenia* (2016), European Language Resources Association (ELRA), pp. 3339–3343.

[27] BUNESCU, R. C., PASCA, M., AND PAŞÇA, M. Using encyclopedic knowledge for named entity disambiguation. In *EACL 2006 - 11th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 5 - 6, 2006, Trento, Italy* (2006), vol. 6, CEUR Workshop Proceedings, pp. 9–16.

[28] BURSTEIN, F., AND GREGOR, S. The Systems Development or Engineering Approach to Research in Information Systems: An Action Research Perspective. In *Proceedings of the 10th Australasian Conference on Information Systems: 1–3 December 1999, Wellington, New Zealand* (1999), Victoria University of Wellington, New Zealand, pp. 122–134.

[29] BUSCALDI, D., ROSSO, P., AND ARNAL, E. S. Using the wordnet ontology in the geoclef geographical information retrieval task. In *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers* (2005), Springer, Springer, Berlin, Heidelberg, pp. 939–946.

[30] CAMPBELL, D. T., AND STANLEY, J. C. *Experimental and Quasi-Experimental Designs for Research*. Ravenio Books, 2015.

[31] CATHRO, W. Metadata: An Overview. In *Proceedings of the Standards Australia Seminar: Matching Discovery and Recovery, Sydney and Melbourne, Australia, August 14-15* (Sydney and Melbourne, Australia, 1997).

[32] CHERVEN, K. *Network graph analysis and visualization with Gephi*. Packt Publishing Ltd, 2013.

[33] CHIARCOS, C., AND FÄTH, C. CoNLL-RDF: Linked corpora done in an NLP-friendly way. In *Language, Data, and Knowledge First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings* (2017), vol. 10318 LNAI, Springer, Cham, pp. 74–88.

[34] CHIARCOS, C., KLIMEK, B., FÄTH, C., DECLERCK, T., AND MCCRAE, J. P. On the Linguistic Linked Open Data Infrastructure. In *Proceedings of the LREC 2020 Workshop IWLTP 2020 – 1st International Workshop on Language Technology Platforms, 16 May 2020, Marseille, France* (Marseille, France, 2020), European Language Resources Association (ELRA), pp. 8–15.

[35] CHINCHOR, N., AND ROBINSON, P. MUC-7 named entity task definition. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998* (1998), pp. 1–21.

[36] CHUNG, Y.-M. M., POTTENGER, W. M., AND SCHATZ, B. R. Automatic subject indexing using an associative neural network. *Proceedings of the ACM International Conference on Digital Libraries, DL98: Digital Libraries '98, Pittsburgh, PA, USA, June 23 - 26, 1998* (1998), 59–68.

[37] CIOTTI, F., AND TOMASI, F. Formal ontologies, linked data, and TEI semantics. *Journal of the Text Encoding Initiative*, 9 (2016).

[38] COMMITTEE ON CATALOGING. Task Force on Metadata. Final Report. Tech. rep., American Library Association, 2000.

[39] COOK, T. D., AND CAMPBELL, D. T. *Quasi-experimentation: Design & Analysis Issues for Field Settings*. Houghton Mifflin, 1979.

[40] COWIE, J., AND LEHNERT, W. Information Extraction. *Communications of the ACM 39*, 1 (1996), 80–91.

[41] Cucerzan, S. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), June 28–30, 2007, Prague, Czech Republic* (Prague, Czech Republic, 2007), Association for Computational Linguistics, pp. 708–716.

[42] Cyganiak, R., Wood, D., and Lanthaler, M. RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation 25 February 2014. `https://www.w3.org/TR/rdf11-concepts/`, 2014. Accessed 30 May 2021.

[43] Damljanovic, D., and Bontcheva, K. Named Entity Disambiguation using Linked Data. In *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings* (2012), Springer-Verlag Berlin Heidelberg, pp. 231–240.

[44] Davison, R. M., Martinsons, M. G., and Kock, N. Principles of canonical action research. *Information Systems Journal 14*, 1 (2004), 65–86.

[45] DCMI Usage Board. DCMI Metadata Terms. `http://dublincore.org/documents/dcmi-terms`, 2012. Accessed 16 September 2021.

[46] DCMI Usage Board. Dublin Core™ Metadata Element Set, Version 1.1: Reference Description. `https://www.dublincore.org/specifications/dublin-core/dces/`, 2012. Accessed 16 September 2021.

[47] Derczynski, L., Maynard, D., Rizzo, G., Van Erp, M., Gorrell, G., Troncy, R., Petrak, J., and Bontcheva, K. Analysis of named entity recognition and linking for tweets. *Information Processing and Management 51*, 2 (3 2015), 32–49.

[48] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, Minnesota, USA, 2019), Association for Computational Linguistics, pp. 4171–4186.

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

[50] Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., and van de Sompel, H. The Europeana Data Model (EDM). In *World Library and Information Congress: 76th IFLA general conference and assembly, August 10-15, 2010, Gothenburg, Sweden* (2010), vol. 10, p. 15.

[51] Doerr, M., Ore, C.-E., and Stead, S. The CIDOC conceptual reference model: a new standard for knowledge sharing. In *Challenges in Conceptual Modelling. Tutorials, posters, panels and industrial contributions at the 26th International Conference on Conceptual Modeling - ER 2007. Auckland, New Zealand, November 5-9, 2007. Proceedings* (2007), vol. 83, Australian Computer Society Inc., pp. 51–56.

[52] Dumont, S. correspSearch – Connecting Scholarly Editions of Letters. *Journal of the Text Encoding Initiative 10* (2016).

[53] Duval, E. Metadata Standards: What, Who & Why. *Journal of Universal Computer Science 7*, 7 (2001), 591–601.

[54] Edelstein, D., Findlen, P., Ceserani, G., Winterer, C., and Coleman, N. Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project Historical Research in a Digital Age. *The American Historical Review 122*, 2 (2017), 400–424.

[55] Elson, D. K., McKeown, K. R., Dames, N., and McKeown, K. R. Extracting Social Networks from Literary Fiction. In *Proceedings of the 48th annual meeting of the association for computational linguistics, ACL 2010, Uppsala, Sweden, July 11-16, 2010* (2010), no. July, Association for Computational Linguistics, pp. 138–147.

[56] Fankhauser, P., Knappen, J., and Teich, E. Exploring and Visualizing Variation in Language Resources. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), May 26 - 31, 2014, Reykjavik, Iceland* (Reykjavik, Iceland, 2014), European Language Resources Association (ELRA), pp. 4125–4128.

[57] Feigenbaum, L., Herman, I., Hongsermeier, T., Neumann, E., and Stephens, S. The semantic web in action. *Scientific American 297*, 6 (2007), 90–97.

[58] Ferragina, P., and Scaiella, U. TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities). In *CIKM '10: Proceedings of the 19th ACM international conference on Information and knowledge management, Toronto, ON, Canada, October 26 - 30, 2010* (2010), Association for Computing Machinery (ACM), p. 1625–1628.

[59] Finkel, J. R., Grenager, T., and Manning, C. D. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), June, 25-30, 2005, University of Michigan, Ann Arbor, Michigan, USA* (2005), Association for Computational Linguistics, pp. 363–370.

[60] Fokkens, A., Soroa, A., Beloki, Z., Ockeloen, N., Rigau, G., Van Hage, W. R., and Vossen, P. NAF and GAF: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation, May 26, 2014, Reykjavik, Iceland* (2014), Association for Computational Linguistics, pp. 9–16.

[61] Fokkens, A., Soroa, A., Beloki, Z., Rigau, G., Hage, W. R. V., and Vossen, P. NAF: the NLP Annotation Format Technical Report NWR-2014-3. Tech. Rep. 3, 2014.

[62] Fokkens, A., ter Braake, S., Ockeloen, N., Vossen, P., Legêne, S., Schreiber, G., and de Boer, V. BiographyNet: Extracting Relations Between People and Events. *Europa baut auf Biographien* (2017), 193–224.

[63] Frosterus, M., Tuominen, J., and Hyvönen, E. Facilitating Re-use of Legal Data in Applications - Finnish Law as a Linked Open Data Service. In *Legal Knowledge and Information Systems - JURIX 2014: The Twenty-Seventh Annual Conference, Jagiellonian University, Krakow, Poland, 10-12 December 2014* (2014), IOS Press, pp. 115–124.

[64] Gasbarra, L., Koho, M., Jokipii, I., Rantala, H., and Hyvönen, E. An Ontology of Finnish Historical Occupations. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (6 2019), A. Poggi, Ed., vol. 11762 LNCS of *CEUR Workshop Proceedings*, pp. 64–68.

[65] GODOY, J., ATKINSON, J., AND RODRIGUEZ, A. Geo-referencing with semi-automatic gazetteer expansion using lexico-syntactical patterns and co-reference analysis. *International Journal of Geographical Information Science 25*, 1 (2011), 149–170.

[66] GOLD, M. K. *Debates in the Digital Humanities*. Debates in the Digital Humanities Series. University of Minnesota Press, 2012.

[67] GORDEA, S., PARAMITA, M. L., AND ISAAC, A. Named entity recommendations to enhance multilingual retrieval in europeana.eu. In *Foundations of Intelligent Systems: 25th International Symposium, ISMIS 2020, Graz, Austria, September 23–25, 2020, Proceedings* (2020), Springer Science and Business Media Deutschland GmbH, pp. 102–112.

[68] GRANDJEAN, M. A social network analysis of Twitter: Mapping the digital humanities community. *Cogent Arts & Humanities 3*, 1 (2016), 1171458.

[69] GRIES, S. T. What is Corpus Linguistics? *Language and Linguistics Compass 3*, 5 (2009), 1225–1241.

[70] GRISHMAN, R., AND SUNDHEIM, B. Message Understanding Conference-6: a brief history. In *16th International Conference on Computational Linguistics, Proceedings of the Conference, COLING 1996, Center for Sprogteknologi, Copenhagen, Denmark, August 5-9, 1996* (1996), vol. 1, Association for Computational Linguistics, p. 466–471.

[71] GROVER, C., TOBIN, R., BYRNE, K., WOOLLARD, M., REID, J., DUNN, S., AND BALL, J. Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368*, 1925 (2010), 3875–3889.

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

[73] GUARINO, N., OBERLE, D., AND STAAB, S. What Is an Ontology? In *Handbook on ontologies*, 2 ed. Springer-Verlag Berlin Heidelberg, 2009, pp. 1–17.

[74] GUHA, R. V., BRICKLEY, D., AND MACBETH, S. Schema.org: evolution of structured data on the web. *Communications of the ACM 59*, 2 (2016), 44–51.

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

[76] HAKOSALO, H., JALAGIN, S., JUNILA, M., AND KURVINEN, H. *Historiallinen elämä - Biografia ja historiantutkimus*. Suomalaisen Kirjallisuuden Seura, Helsinki, Finland, 2014.

[77] HAKULINEN, A., VILKUNA, M., KORHONEN, R., KOIVISTO, V., HEINONEN, T. R., AND ALHO, I. Ison suomen kieliopin verkkoversio. `https://kaino.kotus.fi/visk/etusivu.php`, 2008. Accessed 29 November 2022.

[78] HARRIS, Z. S. Distributional structure. *Word 10*, 2-3 (1954), 146–162.

[79] HAVERINEN, K., NYBLOM, J., VILJANEN, T., LAIPPALA, V., KOHONEN, S., MISSILÄ, A., OJALA, S., SALAKOSKI, T., AND GINTER, F. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation 48*, 3 (2014), 493–531.

[80] Heath, T., and Bizer, C. *Linked Data: Evolving the Web into a Global Data Space*, 1 ed., vol. 1 of *Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool, Palo Alto, CA, USA, 2011.

[81] Hellmann, S., Lehmann, J., and Auer, S. NIF: An ontology-based and linked-data-aware NLP Interchange Format. Tech. rep., 2012. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.232.1215&rep=rep1&type=pdf. Accessed 10 January 2022.

[82] Hellmann, S., Lehmann, J., and Auer, S. Towards an ontology for representing strings. Tech. rep., 2012. http://svn.aksw.org/papers/2012/WWW_NIF/public/string_ontology.pdf. Accessed 10 January 2022.

[83] Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. Integrating NLP using linked data. In *The Semantic Web - ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II* (2013), Springer Berlin Heidelberg, pp. 98–113.

[84] Hevner, A. R., March, S. T., Park, J., and Ram, S. Design science in information systems research. *MIS Quarterly: Management Information Systems 28*, 1 (2004), 75–105.

[85] Heymann, S., and Le Grand, B. Visual analysis of complex networks for business intelligence with gephi. In *2013 17th International Conference on Information Visualisation, IV 2013, July 16 - 18, 2013, London, United Kingdom* (London, United Kingdom, 2013), IEEE Computer Society, pp. 307–312.

[86] Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E., and Coates, C. M. Trading consequences: A case study of combining text mining and visualization to facilitate document exploration. *Digital Scholarship in the Humanities 30*, suppl_1 (2015), i50–i75.

[87] Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., and Rudolph, S. OWL 2 Web Ontology Language Primer (Second Edition). W3C Recommendation 11 December 2012, 2012.

[88] Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. Robust disambiguation of named entities in text. In *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, Edinburgh, United Kingdom, July 27 - 31, 2011* (Stroudsburg, PA, USA, 2011), EMNLP '11, Association for Computational Linguistics, pp. 782–792.

[89] Hu, Y., Janowicz, K., and Prasad, S. Improving Wikipedia-based Place Name Disambiguation in Short Texts Using Structured Data from DBpedia. In *Proceedings of the 8th Workshop on Geographic Information Retrieval, Dallas, TX, USA, November 4 - 7, 2014* (2014), GIR '14, Association for Computing Machinery (ACM), pp. 1–8.

[90] Hussain, S., Muhammad, L. J., and Yakubu, A. Mining social media and DBpedia data using Gephi and R. *Journal of Applied Computer Science & Mathematics 12*, 1 (2018), 14–20.

[91] Hyvönen, E. Semantic Finlex: Finnish law and justice as linked open data. https://seco.cs.aalto.fi/projects/lawlod/en/. Accessed 24 July 2021.

[92] Hyvönen, E. Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery. *Semantic Web 11*, 1 (2020), 187–193.

[93] HYVÖNEN, E. “Sampo” Model and Semantic Portals for Digital Humanities on the Semantic Web. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference, Riga, Latvia, October 21-23, 2020.* (Riga, Latvia, 2020), CEUR Workshop Proceedings, pp. 373–378.

[94] HYVÖNEN, E., HEINO, E., LESKINEN, P., IKKALA, E., KOHO, M., TAMPER, M., TUOMINEN, J., AND MÄKELÄ, E. WarSampo Data Service and Semantic Portal for Publishing Linked Open Data About the Second World War History. In *The Semantic Web. Latest Advances and New Domains - 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Proceedings.* (2016), Springer-Verlag, pp. 758–773.

[95] 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 *Language, Data, and Knowledge First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings* (2017), vol. 10318 LNAI, Springer, Cham, pp. 113–119.

[96] HYVÖNEN, E., LESKINEN, P., TAMPER, M., RANTALA, H., IKKALA, E., TUOMINEN, J., AND KERAVUORI, K. Linked Data – A Paradigm Change for Publishing and Using Biography Collections on the Semantic Web. In *Proceedings of the Third Conference on Biographical Data in a Digital World (BD 2019), Varna, Bulgaria, September, 2019.*

[97] 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 - 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2-6, 2019, Proceedings* (2019), vol. 11503 LNCS, Springer International Publishing, pp. 574–589.

[98] HYVÖNEN, E., LESKINEN, P., TAMPER, M., RANTALA, H., IKKALA, E., TUOMINEN, J., AND KERAVUORI, K. Demonstrating BiographySampo in Solving Digital Humanities Research Problems in Biography and Prosopography. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, Copenhagen, Denmark, March 5-8, 2019.* (Copenhagen, Denmark, 2019), CEUR Workshop Proceedings, pp. 5–7.

[99] HYVÖNEN, E., LESKINEN, P., TAMPER, M., TUOMINEN, J., AND KERAVUORI, K. Semantic National Biography of Finland. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018), Helsinki, Finland, March 7-9, 2018* (Helsinki, Finland, 2018), CEUR Workshop Proceedings, pp. 372–385.

[100] 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 - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers* (2014), vol. 8798, Springer–Verlag, pp. 226–230.

[101] HYVÖNEN, E., TUOMINEN, J., MÄKELÄ, E., DUTRUIT, J., APAJALAHTI, K., HEINO, E., LESKINEN, P., AND IKKALA, E. Second World War on the Semantic Web: The WarSampo Project and Semantic Portal. In *Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located with the 14th International Semantic Web Conference (ISWC-2015), Bethlehem, PA, USA, October 11, 2015.* (2015), S. Villata, J. Z. Pan, and M. Dragoni, Eds., CEUR Workshop Proceedings.

[102] IDE, N., AND ROMARY, L. Outline of the international standard linguistic annotation framework. In *Proceedings of the ACL 2003 workshop on Linguistic annotation: getting the model right, Sapporo, Japan, July 11, 2003* (2003), Association for Computational Linguistics, pp. 1–5.

[103] IDE, N., AND SUDERMAN, K. The Linguistic Annotation Framework: a standard for annotation interchange and merging. *Language Resources and Evaluation 48*, 3 (2014), 395–418.

[104] IDREOS, S., PAPAEMMANOUIL, O., AND CHAUDHURI, S. Overview of data exploration techniques. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria Australia, May 31 - June 4, 2015* (Melbourne, Victoria, Australia, 2015), Association for Computing Machinery (ACM), pp. 277–281.

[105] IKKALA, E., AALTO, T., TUOMINEN, J., AINIALA, T., HYVÖNEN, E., RAUNAMAA, J., UUSITALO, H., AALTO, T., AND AINIALA, T. NameSampo: A linked open data infrastructure and workbench for toponomastic research. In *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Geospatial Humanities, GeoHumanities 2018, Seattle WA, USA, November 6, 2018* (2018), Association for Computing Machinery (ACM), pp. 1–9.

[106] IKKALA, E., HYVÖNEN, E., RANTALA, H., AND KOHO, M. Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. *Semantic Web* (2021), 1–16.

[107] ISAAC, A., AND SUMMERS, E. SKOS Simple Knowledge Organization System Primer. W3C Working Group Note 18 August 2009. `https://www.w3.org/TR/skos-primer/`, 2009. Accessed 01 June 2021.

[108] IVANOVIĆ, M., AND BUDIMAC, Z. An overview of ontologies and data resources in medical domains. *Expert Systems with Applications 41*, 11 (2014), 5158–5166.

[109] JÄNICKE, S., FRANZINI, G., CHEEMA, M. F., AND SCHEUERMANN, G. Visual Text Analysis in Digital Humanities. *Computer Graphics Forum 36*, 6 (2017), 226–250.

[110] JASPER, R., AND USCHOLD, M. A Framework for Understanding and Classifying Ontology Applications. In *Proceedings of the Twelfth Workshop on Knowledge Acquisition, Modeling and Management, Banff, Alberta, Canada, October 16-22* (1999), vol. 99, pp. 16–21.

[111] JATOWT, A., KAWAI, D., AND TANAKA, K. Time-focused analysis of connectivity and popularity of historical persons in Wikipedia. *International Journal on Digital Libraries 20*, 4 (2019), 287–305.

[112] JENSEN, K. E. Linguistics in the digital humanities:(computational) corpus linguistics. *MedieKultur: Journal of media and communication research 30*, 57 (2014), 20–p.

[113] JORDANOUS, A., LAWRENCE, K. F., HEDGES, M., AND TUPMAN, C. Exploring manuscripts: Sharing ancient wisdoms across the semantic web. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, Craiova, Romania, June 13 - 15, 2012* (Craiova, Romania, 2012), Association for Computing Machinery (ACM), pp. 1–12.

[114] JOULIN, A., GRAVE, E., BOJANOWSKI, P., AND MIKOLOV, T. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers* (2017), vol. 2, Association for Computational Linguistics, pp. 427–431.

[115] JURAFSKY, D., AND MARTIN, J. H. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition.*, 2nd editio ed. Pearson Prentice Hall, 2008.

[116] KANERVA, J., GINTER, F., MIEKKA, N., LEINO, A., AND SALAKOSKI, T. Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, October 31 – November 1, 2018, Brussels, Belgium* (Brussels, Belgium, 2018), Association for Computational Linguistics, p. 133–142.

[117] KANSALLISARKISTO. Tuomiokirjahaku. `https://tuomiokirjat.narc.fi/`. Accessed 05 January 2022.

[118] KANSALLISARKISTO. Tuomiokirjat avautuvat vapaaseen käyttöön uudessa verkkopalvelussa. `https://www.epressi.com/tiedotteet/tiede-ja-tutkimus/tuomiokirjat-avautuvat-vapaaseen-kayttoon-uudessa-verkkopalvelussa.html`, 11 2020. Accessed 05 January 2022.

[119] KETTUNEN, K., KUNTTU, T., AND JÄRVELIN, K. To stem or lemmatize a highly inflectional language in a probabilistic IR environment? *Journal of Documentation 61*, 4 (2005), 476–496.

[120] KETTUNEN, K., MÄKELÄ, E., KUOKKALA, J., RUOKOLAINEN, T., AND NIEMI, J. Modern tools for old content-in search of named entities in a finnish ocred historical newspaper collection 1771-1910. In *Lernen, Wissen, Daten, Analysen 2016 : Proceedings of the Conference "Lernen, Wissen, Daten, Analysen", Potsdam, Germany, September 12-14, 2016* (2016), CEUR Workshop Proceedings, pp. 124–135.

[121] KHALILI, A., AUER, S., AND NGOMO, A.-C. N. conTEXT - Lightweight Text Analytics Using Linked Data. In *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014, Proceedings* (2014), Springer, Cham, pp. 628–643.

[122] KIM, S. M., AND CASSIDY, S. Finding Names in Trove: Named Entity Recognition for Australian Historical Newspapers. In *Proceedings of the Australasian Language Technology Association Workshop 2015, December 8–9, 2015, Western Sydney University, Parramatta, Australia* (2015), Association for Computational Linguistics, pp. 57–65.

[123] KLIMEK, B. Proposing an OntoLex-MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models. In *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017) Galway, I* (2017), vol. 1899, CEUR Workshop Proceedings, pp. 68–73.

[124] KLIMEK, B., MCCRAE, J. P., BOSQUE-GIL, J., IONOV, M., TAUBER, J. K., AND CHIARCOS, C. Challenges for the Representations for Morphology in Ontology Lexicons. In *Proceedings of Electronic Lexicography in the 21st Century conference, 1-3 October 2019, Sintra, Portugal* (2019), Lexical Computing CZ s.r.o., pp. 570–591.

[125] KLINGE, M., Ed. *Suomen kansallisbiografia 1–10*. Suomalaisen Kirjallisuuden Seura, Helsinki, Finland, 2007.

[126] KO, A. J., MYERS, B. A., AND CHAU, D. H. A Linguistic Analysis of How People Describe Software Problems. In *Visual Languages and Human-Centric Computing (VL/HCC'06), September 4 - 8, 2006, Brighton, UK* (2006), IEEE, pp. 127–134.

[127] KOHO, M., HEINO, E., AND HYVÖNEN, E. 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 (2016), CEUR Workshop Proceedings.

[128] 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 co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, USA, October 8-12, 2018.* (2018), CEUR Workshop Proceedings, pp. 4–7.

[129] KORENIUS, T., LAURIKKALA, J., JÄRVELIN, K., AND JUHOLA, M. Stemming and lemmatization in the clustering of Finnish text documents. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management, Washington D.C., USA, November 8 - 13, 2004* (2004), Association for Computing Machinery (ACM), pp. 625–633.

[130] KOSKENNIEMI, K. Two-level model for morphological analysis. In *IJCAI'83: Proceedings of the Eighth international joint conference on Artificial intelligence, Karlsruhe, West Germany, August 8 - 12, 1983* (1983), vol. 1, William Kaufmann Inc., pp. 683–685.

[131] KROEGER, A. The Road to BIBFRAME: The Evolution of the Idea of Bibliographic Transition into a Post-MARC Future. *Cataloging & Classification Quarterly 51*, 8 (2013), 873–890.

[132] LAHTI, L., VAARA, V., MARJANEN, J., AND TOLONEN, M. Best practices in bibliographic data science. In *Proceedings of the Research Data And Humanities (RDHUM) 2019 Conference Data, Methods And Tools, University of Oulu, Oulu, Finland, August 14-16, 2019* (2019), Studia humaniora Ouluensia, University of Oulu, pp. 57–65.

[133] LANGMEAD, A., OTIS, J. M., WARREN, C. N., WEINGART, S. B., AND ZILINKSI, L. D. Towards Interoperable Network Ontologies for the Digital Humanities. *International Journal of Humanities and Arts Computing 10*, 1 (2016), 22–35.

[134] LARSON, R. R. Bringing Lives to Light: Biography in Context. Tech. rep., University of Berkeley, 2010.

[135] LATIFI, M., AND SÀNCHEZ-MARRÈ, M. The Use of NLP Interchange Format for Question Answering in Organizations. In *Artificial Intelligence Research and Development - Proceedings of the 16th International Conference of the Catalan Association for Artificial Intelligence, October 23-25, 2013, University of Vic, Vic, Catalonia, Spain* (2013), IOS Press, pp. 235–244.

[136] LAUR, S., ORASMAA, S., SÄRG, D., AND TAMMO, P. EstNLTK 1.6: Remastered Estonian NLP Pipeline. In *Proceedings of The 12th Language Resources and Evaluation Conference, Marseille, France, May 13 - 15, 2020* (2020), European Language Resources Association (ELRA), pp. 7152–7160.

[137] LAUSER, B., AND HOTHO, A. Automatic Multi-label Subject Indexing in a Multilingual Environment. In *Research and Advanced Technology for Digital Libraries, 7th European Conference, ECDL 2003, Trondheim, Norway, August 17-22, 2003. Proceedings.* (2003), Springer, Berlin, Heidelberg, pp. 140–151.

[138] 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, Riga, Latvia, October 21-23, 2020.* (Riga, Latvia, 2020), vol. 2612, CEUR Workshop Proceedings, pp. 284–292.

[139] LESKINEN, P., AND HYVÖNEN, E. Reconciling and Using Historical Person Registers as Linked Open Data in the AcademySampo Knowledge Graph. In *Proceedings of the 20th International Semantic Web Conference (ISWC 2021), October 24-28, 2021, Athens, Greece* (2021), Springer.

[140] LINDQUIST, T., AND LONG, H. How can educational technology facilitate student engagement with online primary sources? A user needs assessment. *Library Hi Tech 29*, 2 (2011), 224–241.

[141] LIU, X., BOLLEN, J., NELSON, M. L., AND DE SOMPEL, H. Co-authorship networks in the digital library research community. *Information processing & management 41*, 6 (2005), 1462–1480.

[142] LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L., AND STOYANOV, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[143] LÖFBERG, L., ARCHER, D., PIAO, S., RAYSON, P., MCENERY, T., VARANTOLA, K., AND JUNTUNEN, J.-P. Porting an English semantic tagger to the Finnish language. In *Proceedings of the Corpus Linguistics 2003 conference, Lancaster University, Lancaster, UK, March 28 - 31, 2003* (Lancaster, UK, 2003), Lancaster University, pp. 457–464.

[144] LOUNELA, M. Exploring morphologically analysed text material. In *Inquiries into words, constraints and contexts: Festschrift in the honour of Kimmo Koskenniemi on his 60th birthday*. CSLI publications, 2005, pp. 259–267.

[145] LUOMA, J., AND PYYSALO, S. Exploring Cross-sentence Contexts for Named Entity Recognition with BERT. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020* (2020), International Committee on Computational Linguistics, pp. 904–914.

[146] MAI, F., GALKE, L., AND SCHERP, A. Using deep learning for title-based semantic subject indexing to reach competitive performance to full-text. In *JCDL '18: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, Fort Worth, TX, USA, June 3 - 7, 2018* (Fort Worth, Texas, USA, 2018), Association for Computing Machinery (ACM), pp. 169–178.

[147] MAIMON, O., AND ROKACH, L. *Data Mining and Knowledge Discovery Handbook*, 2nd ed. Springer Publishing Company, Incorporated, 2010.

[148] 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 - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers* (2014), Springer International Publishing, pp. 424–428.

[149] MÄKELÄ, E. LAS: an integrated language analysis tool for multiple languages. *The Journal of Open Source Software 1*, 6 (2016).

[150] MÄKELÄ, E., LAGUS, K., LAHTI, L., SÄILY, T., TOLONEN, M., HÄMÄLÄINEN, M., KAISLANIEMI, S., AND NEVALAINEN, T. Wrangling with non-standard data. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference, Riga, Latvia, October 21-23, 2020.* (Riga, Latvia, 2020), vol. 2612, CEUR Workshop Proceedings, pp. 81–96.

[151] MÄKELÄ, E., LINDQUIST, T., AND HYVÖNEN, E. CORE - A Contextual Reader based on Linked Data. In *11th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2016, Krakow,*

Poland, July 11-16, 2016, Conference Abstracts (2016), Alliance of Digital Humanities Organizations (ADHO), pp. 267–269.

[152] MANNING, C. D., RAGHAVAN, P., AND SCHÜTZE, H. *Introduction to information retrieval*, vol. 1st ed. Cambridge University Press, New York, NY, USA, 2008.

[153] MANNING, C. D., AND SCHÜTZE, H. *Foundations of statistical natural language processing*, 1st ed. MIT Press, 1999.

[154] MARCH, S. T., AND SMITH, G. F. Design and natural science research on information technology. *Decision Support Systems 15*, 4 (1995), 251–266.

[155] MARTINEZ-RODRIGUEZ, J. L., HOGAN, A., AND LOPEZ-AREVALO, I. Information extraction meets the semantic web: a survey. *Semantic Web 11*, 2 (2020), 255–335.

[156] MAYNARD, D., ROBERTS, I., GREENWOOD, M. A., ROUT, D., AND BONTCHEVA, K. A framework for real-time semantic social media analysis. *Journal of Web Semantics 44* (2017), 75–88.

[157] MCCRAE, J., AGUADO-DE CEA, G., BUITELAAR, P., CIMIANO, P., DECLERCK, T., GÓMEZ-PÉREZ, A., GRACIA, J., HOLLINK, L., MONTIEL-PONSODA, E., SPOHR, D., ET AL. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation 46*, 4 (2012), 701–719.

[158] MCCRAE, J. P., BOSQUE-GIL, J., GRACIA, J., BUITELAAR, P., AND CIMIANO, P. The Ontolex-Lemon model: development and applications. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference, 19-21 September 2017, Leiden, Netherlands* (2017), Lexical Computing CZ s.r.o., pp. 19–21.

[159] MCCRAE, J. P., CHIARCOS, C., BOND, F., CIMIANO, P., DECLERCK, T., DE MELO, G., GRACIA, J., HELLMANN, S., KLIMEK, B., MORAN, S., OSENOVA, P., PAREJA-LORA, A., AND POOL, J. The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), May 23-28, 2016, Portorož, Slovenia* (2016), European Language Resources Association (ELRA), pp. 2435–2441.

[160] MCSWEENEY, P. J. Gephi network statistics. *Google Summer of Code* (2009), 1–8.

[161] MEDELYAN, O., FRANK, E., AND WITTEN, I. H. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL* (2009), Association for Computational Linguistics, pp. 1318–1327.

[162] MEDREK, J., OTTO, C., AND EWERTH, R. Recommending scientific videos based on metadata enrichment using linked open data. In *International Conference on Theory and Practice of Digital Libraries. 22nd International Conference on Theory and Practice of Digital Libraries, September 10-13, 2018, Porto, Portugal* (2018), Springer Cham, pp. 286–292.

[163] MENDES, P. N., JAKOB, M., GARCÍA-SILVA, A., AND BIZER, C. DBpedia Spotlight: Shedding Light on the Web of Documents. In *I-Semantics'11: Proceedings of the 7th international conference on semantic systems, Graz, Austria, September 7 - 9, 2011* (2011), Association for Computing Machinery (ACM), pp. 1–8.

[164] METILLI, D., BARTALESI, V., AND MEGHINI, C. A Wikidata-based tool for building and visualising narratives. *International Journal on Digital Libraries 20*, 4 (2019), 417–432.

[165] MILES, A., AND BECHHOFER, S. SKOS Simple Knowledge Organization System Reference. `https://www.w3.org/TR/2009/REC-skos-reference-20090818/`, 2010. Accessed 05 January 2022.

[166] MIYAKITA, G., LESKINEN, P., AND HYVÖNEN, E. Using Linked Data for Prosopographical Research of Historical Persons: Case U.S. Congress Legislators. 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 II* (2018), vol. 11197 LNCS, Springer International Publishing, pp. 150–162.

[167] MORETTI, F. *Distant Reading*. Verso Books, 2013.

[168] MORETTI, F., AND PIAZZA, A. Graphs, Maps, Trees: Abstract Models for a Literary History. *Modern Language Quarterly 68*, 1 (2007), 132–135.

[169] MORGENSTERN, A., AND AMMINGER, A. Biography as Compilation: How to Encode Georg Nikolaus Nissen's Biographie WA Mozart's (1828) in TEI P5. *Journal of the Text Encoding Initiative*, 11 (2020).

[170] NADEAU, D., AND SEKINE, S. A survey of named entity recognition and classification. *Lingvisticae Investigationes 30*, 1 (2007), 3–26.

[171] NEWMAN, M. *Networks*. Oxford University Press, 2018.

[172] NGUYEN, D. B., HOFFART, J., THEOBALD, M., G., W., NGUYEN, D. B., HOFFART, J., THEOBALD, M., AND WEIKUM, G. AIDA-light: High-Throughput Named-Entity Disambiguation. In *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014* (2014), CEUR Workshop Proceedings.

[173] NIVRE, J., DE MARNEFFE, M. C., GINTER, F., GOLDBERG, Y., HAJIČ, J., MANNING, C. D., MCDONALD, R., PETROV, S., PYYSALO, S., SILVEIRA, N., TSARFATY, R., AND ZEMAN, D. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), May 23-28, 2016, Portorož, Slovenia* (2016), European Language Resources Association (ELRA), pp. 1659–1666.

[174] OCKELOEN, N., FOKKENS, A., TER BRAAKE, S., VOSSEN, P., DE BOER, V., SCHREIBER, G., AND LEGÊNE, S. BiographyNet: Managing Provenance at Multiple Levels and from Different Perspectives. In *Proceedings of the 3rd International Workshop on Linked Science 2013 - Supporting Reproducibility, Scientific Investigations and Experiments (LISC2013) In conjunction with the 12th International Semantic Web Conference 2013 (ISWC 2013), Sydney, Australia,* (2013), CEUR Workshop Proceedings, pp. 59–71.

[175] OKSANEN, A., TAMPER, M., TUOMINEN, J., MÄKELÄ, E., HIETANEN, A., AND HYVÖNEN, E. Semantic Finlex: Transforming, publishing, and using finnish legislation and case law as linked open data on the web. In *Knowledge of the Law in the Big Data Age*, vol. 317. IOS Press, Amsterdam, Netherlands, 2019, pp. 212–228.

[176] OTTE, E., AND ROUSSEAU, R. Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science 28*, 6 (2002), 441–453.

[177] PAIK, J. H. A Novel TF-IDF Weighting Scheme for Effective Ranking. In *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland, July 28 - August 01, 2013* (2013), SIGIR '13, Association for Computing Machinery (ACM), pp. 343–352.

[178] PARSONS, S. *A Semantic Web Primer*, 3rd ed., vol. 24. MIT Press, Cambridge, MA, USA, 2009.

[179] PATTUELLI, M. C., MILLER, M., LANGE, L., AND THORSEN, H. Linked Jazz 52nd Street: A LOD Crowdsourcing Tool to Reveal Connections among Jazz Artists. In *8th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2013, Lincoln, NE, USA, July 16-19, 2013, Conference Abstracts* (2013), Alliance of Digital Humanities Organizations (ADHO), pp. 337–339.

[180] PAULHEIM, H. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. *Semantic web 8*, 3 (2017), 489–508.

[181] PEFFERS, K., TUUNANEN, T., ROTHENBERGER, M. A., AND CHATTERJEE, S. A design science research methodology for information systems research. *Journal of Management Information Systems 24*, 3 (2007), 45–77.

[182] PENG, N., FERRARO, F., YU, M., ANDREWS, N., DEYOUNG, J., THOMAS, M., GORMLEY, M. R., WOLFE, T., HARMAN, C., VAN DURME, B., AND OTHERS. A concrete Chinese NLP pipeline. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, HLT-NAACL 2015, May 31 - June 5, 2015, Denver, Colorado, USA* (2015), Association for Computational Linguistics, pp. 86–90.

[183] PETRAS, V., HILL, T., STILLER, J., AND GÄDE, M. Europeana - a search engine for digitised cultural heritage material. *Datenbank-Spektrum 17*, 1 (2017), 41–46.

[184] PHIPPS, J., DUNSIRE, G., AND HILLMANN, D. Building a Platform to Manage RDA Vocabularies and Data for an International, Linked Data World. *Journal of Library Metadata 15*, 3-4 (2015), 252–264.

[185] PIATESKI, G., AND FRAWLEY, W. *Knowledge Discovery in Databases*. MIT Press, Cambridge, MA, USA, 1991.

[186] PICCINNO, F., AND FERRAGINA, P. From Tagme to WAT: A new entity annotator. In *ERD'14, Proceedings of the First ACM International Workshop on Entity Recognition & Disambiguation, July 11, 2014, Gold Coast, Queensland, Australia* (2014), Association for Computing Machinery (ACM), pp. 55–61.

[187] PIRINEN, T. A. Omorfi—Free and open source morphological lexical database for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015), May 11 - 13, 2015, Vilnius, Lithuania* (2015), Linköping University Electronic Press, Sweden, pp. 313–316.

[188] PIRJO MIKKONEN, S. P. *Sukunimet*. Otavan kirjapaino Oy, 2000.

[189] POULIQUEN, B., STEINBERGER, R., AND IGNAT, C. Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. In *Proceedings of the Workshop 'Ontologies and Information Extraction' at the Summer School 'The Semantic Web and Language Technology - Its Potential and Practicalities' (EUROLAN 2003), 28 July - 8 August, 2003, Bucharest, Romania* (2006), pp. 8–28.

[190] POWELL, A., NILSSON, M., NAEVE, A., JOHNSTON, P., AND BAKER, T. DCMI Abstract Model. https://www.dublincore.org/specifications/dublin-core/abstract-model/2007-06-04/, 2007.

[191] PRABHU, Y., AND VARMA, M. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *KDD '14: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, August 24 - 27, 2014, New York, NY, USA* (2014), Association for Computing Machinery (ACM), pp. 263–272.

[192] PRESUTTI, V., DRAICCHIO, F., AND GANGEMI, A. Knowledge extraction based on discourse representation theory and linguistic frames. In *Knowledge Engineering and Knowledge Management: 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012, Proceedings* (2012), Association for Computing Machinery (ACM), pp. 114–129.

[193] PYYSALO, S., AND GINTER, F. Collaborative development of annotation guidelines with application to Universal Dependencies. In *The Fifth Swedish Language Technology Conference, SLTC 2014, November 13-14, 2014, Uppsala, Sweden* (2014), SLTC.

[194] QI, P., ZHANG, Y., ZHANG, Y., BOLTON, J., AND MANNING, C. D. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Online, 2020), Association for Computational Linguistics, pp. 101–108.

[195] RAVENEK, W., VAN DEN HEUVEL, C., AND GERRITSEN, G. The ePistolarium: origins and techniques. In *CLARIN in the low countries*. Ubiquity Press, 2017, pp. 317–324.

[196] RIETVELD, L., AND HOEKSTRA, R. The YASGUI family of SPARQL clients 1. *Semantic Web 8*, 3 (12 2017), 373–383.

[197] RILEY, J. *Understanding Metadata: What is Metadata, and what is it For?* National Information Standards Organization, 2017.

[198] RILOFF, E., AND JONES, R. Learning Dictionaries for Information Extraction by Multi-level Bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99), July 18 - 22, 1999, Orlando,* (1999), AAAI Press, pp. 474–479.

[199] ROSENTHAL, G. Biographical Research. In *Qualitative Research Practice*, Understanding social research. SAGE, 2004, pp. 49–65.

[200] ROSPOCHER, M., VAN ERP, M., VOSSEN, P., FOKKENS, A., ALDABE, I., RIGAU, G., SOROA, A., PLOEGER, T., AND BOGAARD, T. Building event-centric knowledge graphs from news. *Journal of Web Semantics: Knowledge Technologies 37* (2016), 132–151.

[201] RUOKOLAINEN, T., KAUPPINEN, P., SILFVERBERG, M., AND LINDÉN, K. A Finnish news corpus for named entity recognition. *Language Resources and Evaluation 54*, 1 (8 2020), 247–272.

[202] RUOKOLAINEN, T., AND KETTUNEN, K. À la recherche du nom perdu– searching for named entities with Stanford NER in a Finnish historical newspaper and journal collection. In *13th IAPR International Workshop on Document Analysis Systems, DAS 2018, Vienna, Austria, April 24-27, 2018* (2018), IEEE Computer Society.

[203] SCHILING, V. Transforming Library Metadata into Linked Library Data. *American Library Association* (2012).

[204] SCHREIBMAN, R. S. S., AND UNSWORTH, J., Eds. *A Companion to Digital Humanities*. Blackwell Publishing Ltd, 2004.

[205] 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 (2015), 443–460.

[206] SILVERMAN, H., AND RUGGLES, D. F. Cultural Heritage and Human Rights. In *Cultural Heritage and Human Rights*. Springer-Verlag New York, 2007, pp. 3–29.

[207] SIMON, R., BARKER, E., ISAKSEN, L., AND DE SOTO CAÑAMARES, P. Linking early geospatial documents, one place at a time: annotation of geographic documents with Recogito. *e-Perimetron 10*, 2 (2015), 49–59.

[208] SIMON, R., BARKER, E., ISAKSEN, L., AND DE SOTO CAÑAMARES, P. Linked data annotation without the pointy brackets: Introducing Recogito 2. *Journal of Map & Geography Libraries 13*, 1 (2017), 111–132.

[209] SINCLAIR, S., AND ROCKWELL, G. Voyant Tools. https://voyant-tools.org/docs/#!/guide/about, 2016. Accessed 01 October 2021.

[210] SINGH, K., BOTH, A., DIEFENBACH, D., AND SHEKARPOUR, S. Towards a message-driven vocabulary for promoting the interoperability of question answering systems. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC 2016), February 4-6, 2016, Laguna Hills, CA, USA* (2016), IEEE Computer Society, pp. 386–389.

[211] SINKKILÄ, R., SUOMINEN, O., AND HYVÖNEN, E. Automatic semantic subject indexing of web documents in highly inflected languages. In *The Semantic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29-June 2, 2011, Proceedings, Part I* (2011), Springer-Verlag Berlin Heidelberg, pp. 215–229.

[212] SMALL, H. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science 24*, 4 (1973), 265–269.

[213] SONG, M., AND CHAMBERS, T. Text Mining with the Stanford CoreNLP. In *Measuring Scholarly Impact: Methods and Practice*. Springer, Cham, 2014, pp. 215–234.

[214] SOTIROVA, K., PENEVA, J., IVANOV, S., DONEVA, R., AND DOBREVA, M. Digitization of Cultural Heritage - Standards, Institutions, Initiatives. In *Access to digital cultural heritage: Innovative applications of automated metadata generation*. Plovdiv University Publishing House, 2012, pp. 23–68.

[215] STAAB, S., MAEDCHE, A., AND HANDSCHUH, S. An Annotation Framework for the Semantic Web. In *Proceedings of the First International Workshop on Multimedia Annotation, MMA-2001, January 30-31, 2001, Tokyo, Japan* (2001).

[216] STAAB, S., AND STUDER, R. *Handbook on Ontologies*, 2nd ed. Springer Publishing Company, Incorporated, 2009.

[217] STUDER, R., BENJAMINS, V. R., AND FENSEL, D. Knowledge Engineering: Principles and Methods. *Data & Knowledge Engineering 25*, 1-2 (3 1998), 161–197.

[218] SUOMALAISEN KIRJALLISUUDEN SEURA. Artikkelien rakenne ja lähteet. https://kansallisbiografia.fi/kansallisbiografia/artikkelien-rakenne-ja-lahteet#lahteet. Accessed 05 January 2022.

[219] SUOMALAISEN KIRJALLISUUDEN SEURA. Ohjeita Kansallisbiografian kirjoittajille. http://kansallisbiografia.fi/kansallisbiografia/ohjeita-kirjoittajille. Accessed 05 January 2021.

[220] SUOMINEN, O. *Methods for Building Semantic Portals*. PhD thesis, Aalto University, 2013.

[221] SUOMINEN, O. Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly 29*, 1 (7 2019), 1.

[222] SUOMINEN, O., AND HYVÖNEN, N. From MARC silos to Linked Data silos? *o-bib. Das offene Bibliotheksjournal 4*, 2 (2017), 1–13.

[223] TAMPER, M., OKSANEN, A., TUOMINEN, J., HIETANEN, A., AND HYVÖNEN, E. Automatic Annotation Service APPI: Named Entity Linking in Legal Domain. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15* (2020), Springer International Publishing, pp. 110–114.

[224] TAPANAINEN, P., AND JÄRVINEN, T. A Non-projective Dependency Parser. In *Proceedings of the Fifth Conference on Applied Natural Language Processing, ANLC 1997, March 31 - April 3, 1997, Washington, DC, USA* (1997), Association for Computational Linguistics, pp. 64–71.

[225] TEI CONSORTIUM, E. TEI P5: Guidelines for electronic text encoding and exchange. http://www.tei-c.org/Guidelines/P5/, 2015.

[226] THE ASSOCIATION FOR MILITARY HISTORY IN FINLAND. Kansa Taisteli lehdet 1957 – 1986. http://www.sshs.fi/sitenews/view/-/nid/92/ngid/1, 2014.

[227] THE W3C SPARQL WORKING GROUP. SPARQL 1.1 Overview. W3C Recommendation 21 March 2013. https://www.w3.org/TR/sparql11-overview/, 2013. Accessed 3 May 2022.

[228] TILLETT, B. What is FRBR? A conceptual model for the bibliographic universe. *Australian Library Journal 54*, 1 (2005), 24–30.

[229] TILLETT, B. Keeping libraries relevant in the Semantic Web with resource description and access (RDA). *Serials 24*, 3 (2011), 266–272.

[230] TJONG KIM SANG, E. F., AND DE MEULDER, F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 – June 1, 2003.* (2003), Association for Computational Linguistics, pp. 142–147.

[231] TOGNINI-BONELLI, E. *Corpus Linguistics at Work*. John Benjamins Publishing Company, April 2001.

[232] TUNKELANG, D. *Faceted Search*. Morgan & Claypool, Palo Alto, CA, USA, 2009.

[233] TUOMINEN, J., LAURENNE, N., KOHO, M., AND HYVÖNEN, E. The Birds of the World Ontology AVIO. In *The Semantic Web: ESWC 2013 Satellite Events - ESWC 2013 Satellite Events, Montpellier, France, May 26-30, 2013, Revised Selected Papers.* (2013), vol. 7955 LNCS, Springer-Verlag Berlin Heidelberg, pp. 300–301.

[234] Tuominen, J., Mäkelä, E., Hyvönen, E., Bosse, A., Lewis, M., and Hotson, H. Reassembling the republic of letters - A linked data approach. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018), Helsinki, Finland, March 7-9, 2018* (2018), CEUR Workshop Proceedings, pp. 76–88.

[235] Turco, M. L., Calvano, M., Giovannini, E. C., Lo Turco, M., Calvano, M., and Giovannini, E. C. DATA MODELING for MUSEUM COLLECTIONS. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-2/W9, 2019 8th Intl. Workshop 3D-ARCH "3D Virtual Reconstruction and Visualization of Complex Architectures", 6–8 February 2019, Bergamo, Italy* (2019), vol. XLII-2/W9, Copernicus GmbH, pp. 433–440.

[236] Vajjala, S., and Balasubramaniam, R. What do we really know about state of the art NER? In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (Marseille, France, 2022), European Language Resources Association (ELRA), pp. 5983–5993.

[237] 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), collocated with CHI 2008, Florence, Italy, April 5, 2008.* (2008), CEUR Workshop Proceedings.

[238] Vanhoutte, E. An Introduction to the TEI and the TEI Consortium. *Literary and linguistic computing 19*, 1 (2004), 9–16.

[239] Vecco, M. A definition of cultural heritage: From the tangible to the intangible. *Journal of Cultural Heritage 11*, 3 (2010), 321–324.

[240] Verboven, K., Carlier, M., and Dumolyn, J. A short manual to the art of prosopography. In *Prosopography approaches and applications. A handbook*. Unit for Prosopographical Research (Linacre College), Hockley, Essex, 2007, ch. Nature of, pp. 35–70.

[241] Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. Multilingual is not enough: BERT for Finnish. *CoRR abs/1912.0* (2019).

[242] Volz, R., Kleb, J., and Mueller, W. Towards Ontology-based Disambiguation of Geographical Identifiers. In *Proceedings of the Workshop on Entity-Centric Approaches to Information and Knowledge Management on the Web co-located with WWW2007, Banff, Canada, May 8, 2007* (2007), CEUR Workshop Proceedings.

[243] Vossen, P., Agerri, R., Aldabe, I., Cybulska, A., van Erp, M., Fokkens, A., Laparra, E., Minard, A.-L., Palmero Aprosio, A., Rigau, G., Rospocher, M., and Segers, R. NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems 110* (2016), 60–85.

[244] Vrandečić, D., and Krötzsch, M. Wikidata: a free collaborative knowledgebase. *Communications of the ACM 57*, 10 (2014), 78–85.

[245] W3C, Brickley, D., and Guha, R. V. RDF Schema 1.1. W3C Recommendation 25 February 2014. `https://www.w3.org/TR/rdf-schema/`, 2014. Accessed 01 Jun 2021.

[246] Warren, C. N. Historiography's Two Voices: Data Infrastructure and History at Scale in the Oxford Dictionary of National Biography (ODNB). *Journal of Cultural Analytics 1*, 2 (2018), 1–31.

[247] WARREN, C. N., WARREN, C. N., SHORE, D., OTIS, J., WANG, L., FINEGOLD, M., AND SHALIZI, C. Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks. *Digital Humanities Quarterly 010*, 3 (2016), 1–16.

[248] WENTLAND, W., KNOPP, J., SILBERER, C., AND HARTUNG, M. Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), May 28 - 30, 2008, Marrakech, Morocco* (5 2008), European Language Resources Association (ELRA).

[249] XU, A., HESS, K., AND AKERMAN, L. From MARC to BIBFRAME 2.0: Crosswalks. *Cataloging & Classification Quarterly 56*, 2-3 (2018), 224–250.

[250] YAMASHITA, T., AND MATSUMOTO, Y. Language independent morphological analysis. In *Proceedings of the sixth conference on Applied natural language processing, ANLP 2000, April 29 - May 4, 2000, Seattle, WA, USA* (2000), Association for Computational Linguistics, pp. 232–238.

[251] YEN, I. E.-H., HUANG, X., RAVIKUMAR, P., ZHONG, K., AND DHILLON, I. PD-Sparse : A Primal and Dual Sparse Approach to Extreme Multiclass and Multilabel Classification. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016* (2016), JMLR: Workshop and Conference Proceedings, pp. 3069–3077.

[252] YIMAM, S. M., BIEMANN, C., ECKART DE CASTILHO, R., GUREVYCH, I., DE CASTILHO, R., AND GUREVYCH, I. Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, June 23-24, 2014, Baltimore, MD, USA.* (6 2015), 91–96.

[253] ZHU, G., AND IGLESIAS, C. A. Exploiting semantic similarity for named entity disambiguation in knowledge graphs. *Expert Systems with Applications 101* (2018), 8–24.

The digitization of Cultural Heritage collections has enabled the use of computational methods such as Natural Language Processing (NLP) on textual collections. These methods have been used widely in Digital Humanities (DH) to study digitized contents with automated processes. The Semantic Web and linked data technologies have been applied to describe documents and their metadata, such as author, title, or manually assigned keywords. Other information about the content is often scarce and finding text related to an actor can be laborious. This thesis studies models, methods, and tools for transforming and enriching document collections to linked data. Linked data technology can be used to link texts based on their metadata and information extracted from text, such as mentioned actors. This thesis aims to study how the NLP methods and linked data can be used to study digitized document collections, such as biographies. The thesis presents a toolkit that can be used to model, transform, and enrich biographical text document collections to linked data to improve their information retrieval and interoperability internally and with other collections. The data model describing text document collection's content and features creates a basis for building linked-data-based intelligent services, such as network or linguistic analysis. This approach can be also used to evaluate the quality of text document collections for DH research.

Aalto-DT 19/2023

9 789526 411507

ISBN 978-952-64-1150-7 (printed)
ISBN 978-952-64-1151-4 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Computer Science
www.aalto.fi