(DRAFT)

Modeling and Using Biographical Linked Data for Prosopographical Data Analysis

Petri Leskinen

February 8, 2024

Preface

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

I would like the thank my supervisor, Professor Eero Hyvönen who enabled the project and my advisor, Doctor Jouni Tuominen as well as all the co-authors of the articles Erkki Heino, Esko Ikkala, Mikko Koho, Goki Miyakita, Eetu Mäkelä, Heikki Rantala, and Minna Tamper for collaborations. I also thank Kirsi Keravuori, Risto Valjus, and Susanna Ånäs for co-operation with BiographySampo and Mikko Kivelä and Javier Ureña-Carrion for co-operation with LetterSampo. Thanks to Yrjö Kotivuori and Veli-Matti Autio for their seminal work in creating the original databases used in our work AcademySampo, and for making the data openly available. Thanks to Stefan Dumont for providing the correspSearch data for our use, and Charles van den Heuvel and Dirk van Miert for discussions related to CKCC. I would also like to thank all the other past and current members of SeCo and HELDIG as well as the members in the projects In/Tangible European Heritage (InTaVia) and Constellations of Correspondence (CoCo).

Part of my research was funded by the Severi project[1], funded mainly by Business Finland. Developing the project BiographySampo is also part of the Open Science and Research Programme[2], funded by the Ministry of Education and Culture of Finland. The work was also part of the EU project InTaVia: In/Tangible European Heritage[3], and is related to the EU COST action Nexus xLinguarum[4] on linguistic data science. CSC – IT Center for Science[5] provided computational resources. Furthermore, parts of the work are related to the projects *FIN-CLARIAH* and *ARIADNEplus*.

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

[2] https://openscience.fi

[3] https://intavia.eu/

[4] https://nexuslinguarum.eu/the-action

[5] https://www.csc.fi/en/home

Finally, I will thank my dog *Vilperi* for taking me out for long walks into the woods!

Espoo, February 8, 2024,

Petri Leskinen

Abstract

Biographical data is used for identifying people, groups, and organizations and for representing information about them. Biographical data describes life stories of people with the aim of getting a better understanding of their personality and actions. The underlying texts can also be used for data analysis and distant reading once the documents are provided in a machine-readable format. Prosopographical analysis delves into the life stories of individuals within a defined group to identify shared characteristics and patterns.

This dissertation presents a comprehensive framework for managing and analyzing biographical data in Digital Humanities research. It includes data models, methods, and applications that enrich biographical content with links and reasoning to enhance user experience. Furthermore, the framework provides versatile tools for both individual biographical research and prosopographical research on groups of people.

Linked Data together with event-based data model schemas are used in the published datasets to achieve the interoperability of heterogeneous datasets regarding historical people. Events are used as the glue combining information from various sources. The event-based modeling enables depicting historical narratives as data, which can be further enriched with the events of individual people and organizations.

The research included in this dissertation has been carried out in multiple research projects concentrating on biographical data: WarSampo *(2015–),* BiographySampo *(2018–2021),* Norssi High School Alumni *(2017),* AcademySampo *(2019–2021),* LetterSampo *(2020–2022), and* ParliamentSampo *(2021–2023). The data publications, online portals, and published articles with analysis are represented as the results of the work accomplished for this thesis. Besides, this thesis tackles the practices of creating, modeling, and publishing Linked Data, as well as on analyzing this biographical and prosopographical data by the means of network and data analysis.*

Keywords: Biographical Data, Data Analysis, Digital Humanities, Linked Open Data, Network Analysis, Prosopography, Semantic Web

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 Petri Leskinen, Mikko Koho, Erkki Heino, Minna Tamper, Esko Ikkala, Jouni Tuominen, Eetu Mäkelä, and Eero Hyvönen. Modeling and Using an Actor Ontology of Second World War Military Units and Personnel. In *The Semantic Web – ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part II*, Claudia d'Amato, Miriam Fernandez, Valentina Tamma, Freddy Lecue, Philippe Cudré-Mauroux, Juan Sequeda, Christoph Lange, and Jeff Heflin (editors), Information Systems and Applications, incl. Internet/Web, and HCI, volume 10588, pages 280–296, ISBN 9783319682037, Springer, Cham, October 2017, online https://doi.org/10.1007/978-3-319-68204-4_27.

II Petri Leskinen, Jouni Tuominen, Erkki Heino, and Eero Hyvönen. An Ontology and Data Infrastructure for Publishing and Using Biographical Linked Data. In *Proceedings of the Workshop on Humanities in the Semantic Web (WHiSe II)*, Alessandro Adamou, Enrico Daga, Leif Isaksen (editors), CEUR Workshop Proceedings, pages 15-26, Vienna, Austria, October, 2017, online `https://ceur-ws.org/Vol-2014/paper-02.pdf`.

III Petri Leskinen, Eero Hyvönen, and Jouni Tuominen. Analyzing and Visualizing Prosopographical Linked Data Based on Short Biographies. In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)*, Antske Fokkens, Serge ter Braake, Ronald Sluijter, Paul Arthur, Eveline Wandl-Vogt (editors), pages 39–44, CEUR Workshop Proceedings, Linz, Austria, June 2018, online `http://ceur-ws.org/Vol-2119/paper7.pdf`.

IV Petri Leskinen, Goki Miyakita, Mikko Koho, and Eero Hyvönen. Combining Faceted Search with Data-analytic Visualizations on Top of a

SPARQL Endpoint. In *Proceedings of VOILA 2018, Monterey, California. CEUR Workshop Proceedings, Vol. 2187*, Valentina Ivanova, Patrick Lambrix, Steffen Lohmann, Catia Pesquita (editors), Monterey, CA, USA, August 2018, online https://ceur-ws.org/Vol-2187/paper5.pdf .

V Eero Hyvönen, Petri Leskinen, Minna Tamper, Heikki Rantala, Esko Ikkala, Jouni Tuominen, and Kirsi Keravuori. 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*, Pascal Hitzler, Miriam Fernández, Krzysztof Janowicz, Amrapali Zaveri, Alasdair J. G. Gray, Vanessa Lopez, Armin Haller, and Karl Hammar (editors), Lecture Notes in Computer Science, volume 11503, pages 574–589, Springer-Verlag, June 2019, online https://doi.org/10.1007/978-3-030-21348-0_37.

VI Petri Leskinen and Eero Hyvönen. Extracting Genealogical Networks of Linked Data from Biographical Texts. In *The Semantic Web: ESWC 2019 Satellite Events*, Pascal Hitzler, Sabrina Kirrane, Olaf Hartig, Victor de Boer, Maria-Esther Vidal, Maria Maleshkova, Stefan Schlobach, Karl Hammar, Nelia Lasierra, Steffen Stadtmüller, Katja Hose, Ruben Verborgh (editors), June 2019, pages 121–125, Springer, ISBN 978-3-030-32327-1, online https://doi.org/10.1007/978-3-030-32327-1_24.

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, December 2022, ISSN 1570-0844 (P), DOI 10.3233/SQ-222887, online https://doi.org/10.3233/SW-222887 .

VIII Petri Leskinen and Eero Hyvönen. Linked Open Data Service about Historical Finnish Academic People in 1640–1899. In *DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, Riga, Latvia, Sanita Reinsone, Inguna Skadiņa, Anda Baklāne, Jānis Daugavietis (editors), pages 284–292, CEUR Workshop Proceedings, vol. 2612. October 2020. online https://ceur-ws.org/Vol-2612/short14.pdf.

IX Petri Leskinen and Eero Hyvönen. Reconciling and Using Historical Person Registers as Linked Open Data in the AcademySampo Portal and Data Service. *ISWC2021*, Andreas Hotho, Eva Blomqvist, Stefan

Dietze, Achille Fokoue, Ying Ding, Payam Barnaghi, Armin Haller, Mauro Dragoni, Harith Alani (editors), pages 714—730, Springer, October 2021. online https://doi.org/10.1007/978-3-030-88361-4_42.

X Minna Tamper, Petri Leskinen, Jouni Tuominen, and Eero Hyvönen. Modeling and Publishing Finnish Person Names as a Linked Open Data Ontology. *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, June 2020, pages 3–14, ISSN 1613-0073, online `http://ceur-ws.org/Vol-2695/paper1.pdf` .

XI Petri Leskinen, Javier Ureña-Carrion, Jouni Tuominen, Mikko Kivelä, and Eero Hyvönen. Knowledge Graphs and Data Services for Studying Historical Epistolary Data in Network Science on the Semantic Web. *Submitted for review*, Semantic Web Journal .

XII Petri Leskinen, Eero Hyvönen, and Jouni Tuominen. Members of Parliament in Finland Knowledge Graph and its Linked Open Data Service. *Further with Knowledge Graphs. Proceedings of the 17th International Conference on Semantic Systems, 6-9 September 2021, Amsterdam, The Netherlands.*, Mehwish Alam, Paul Groth, Victor de Boer, Tassilo Pellegrini, Harshvardhan J. Pandit, Elena Montiel, Víctor Rodríguez Doncel, Barbara McGillivray, Albert Meroño-Peñuela (editors), IOS Press, pages 255–269, DOI 10.3233/SSW210049, online https://doi.org/10.3233/SW-210049 .

Author's Contribution

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

DC (Doctoral Candidate) is the lead author of the publication. DC wrote about 70% of the article. DC is the primary developer of the actor ontology data model in co-operation with A, B, C, D, E, F, and G (other authors based on the order in the article). DC is the main producer of the corresponding actor dataset described in chapters 2–4 and portal pages related to actors in chapter 5 including the data diagrams and tables. Subchapter 5.3 was written by C (around 10% of the article), and subchapter 5.4 by B (less than 5% of the article). Casualties data was produced by A and B, Magazine data by C, and the place ontology by D. Chapters Abstract, Introduction, and Discussion (15% of the article) were written in co-operation between all authors. This article is based on the Master's Thesis of DC written in Finnish, and the idea for the paper was proposed by G.

Publication II: "An Ontology and Data Infrastructure for Publishing and Using Biographical Linked Data"

DC the lead author of the publication and wrote about 90% of the article. The person ontology data model was developed by DC and B, as well as the online portal, and the adaptation of Bio CRM in co-operation with A who is one of the main developers of Bio CRM. Chapters Abstract, Introduction, and Discussion were written in co-operation between all authors. The idea for this paper was proposed by C.

Publication III: "Analyzing and Visualizing Prosopographical Linked Data Based on Short Biographies"

DC is the lead author of the publication and wrote about 85% of the article. DC is the primary developer of the represented data publications and online visualizations. Abstract, Introduction, and Discussion were written in co-operation between all authors. The idea for this paper was proposed by A.

Publication IV: "Combining Faceted Search with Data-analytic Visualizations on Top of a SPARQL Endpoint"

DC is the lead author of the publication. DC is one of the developers of the represented online visualizations, and wrote about 25% of the article regarding the datasets *Norssit* and *Semantic National Biography of Finland*. B wrote the chapters introducing the dataset *WarSampo*, and A the chapters introducing *U.S. Congress Prosopographer*. Abstract, Introduction, and Discussion were written in co-operation between all the authors. The idea for this paper was proposed by A.

Publication V: "BiographySampo – Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research"

DC contributed to the writing of the publication as a co-author and wrote about 25% of the article. DC wrote the first half of the chapter *Knowledge Extraction* introducing pattern-based information extraction. The other half was written by B. DC also wrote the subchapters introducing data model, actor ontologies, networks, and the visualizations. Subchapter *6. Relations* was written by C, *2. Places* by D, and *7. Language* by B. Abstract, Introduction, and Discussion were written in co-operation between all authors. The idea for this paper was proposed by A.

Publication VI: "Extracting Genealogical Networks of Linked Data from Biographical Texts"

DC is the lead author of the publication, and wrote about 85% of the article. DC is the main developer of the data conversion process and generated dataset. DC evaluated the linking quality of the process represented in the publication. Abstract and Introduction were written in co-operation between the authors. The idea for this paper was proposed by DC.

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

DC contributed to the writing as a co-author and wrote 30% of the article. DC created statistics and visualizations related to relatives, events, maps, and network metrics while A for authors, text analysis, and network's reference analysis. All the results were evaluated in collaboration with C and D. Abstract, Introduction, and Discussion were written in co-operation between all authors. The idea for the article was proposed by B.

Publication VIII: "Linked Open Data Service about Historical Finnish Academic People in 1640–1899"

DC is the lead author of the publication, and wrote about 85% of the article. DC is the primary designer of the represented data transformation process, data analysis results, and web application. Abstract, Introduction, and the chapter Related Work and Discussion were written in co-operation between the authors. The idea for the article was proposed by A and DC.

Publication IX: "Reconciling and Using Historical Person Registers as Linked Open Data in the AcademySampo Portal and Data Service"

DC is the lead author of the publication, and wrote about 90% of the article. DC is the primary designer of the represented data process. Abstract, Introduction, and Discussion were written in co-operation between the authors. The idea for the article was proposed by DC.

Publication X: "Modeling and Publishing Finnish Person Names as a Linked Open Data Ontology"

DC contributed to the writing of the publication as a co-author, and wrote about 10% of the article. DC is the developer of the statistical calculations for *the Gender Identification Service*, e.g., a system to infer a person's gender based on the given names. DC also transformed data from other Sampo systems into the HENKO ontology and data service. A acted as the main designer and developer of the data service and harvested and converted the Digital Agency dataset of names. The idea for the paper was proposed by A and C.

Publication XI: "Knowledge Graphs and Data Services for Studying Historical Epistolary Data in Network Science on the Semantic Web"

DC is the lead author of the publication, and wrote about 60% of the article including chapters 3 and 4 and the represented data visualizations. The data model was designed in co-operation with B and D. The rest of the article was written in co-operation between all the authors. The idea for the paper was proposed by D.

Publication XII: "Members of Parliament in Finland Knowledge Graph and its Linked Open Data Service"

DC is the lead author of the publication, and wrote about 65% of the article. DC is also the primary implementer of the represented data transformation process. The data model was designed in co-operation with A and B as well as writing the rest of the article (35%) including Abstract, Introduction, and Discussion.

Furthermore, all the related projects have been carried out in co-operation between multiple authors. In addition to the aforementioned publications, the thesis contains references to related work by the author concerning Norssi High School Alumni [112], WarSampo [119, 131, 132], AcademySampo [144, 147, 145], BiographySampo [111, 113, 114, 115, 195, 213, 215], and ParliamentSampo [116, 214]. Finally, a general overview of biographical data publications including user statistics is provided in [109].

Abbreviations

API Application Programming Interface

BIO Vocabulary for Biographical Information

CIDOC International Committee for Documentation

CIDOC CRM CIDOC Conceptual Reference Model

CH Culture Heritage

CKCC Circulation of Knowledge, A Web-based Humanities' Collaboratory on Correspondences

CMI Correspondence Metadata Interchange Format

CoLab Google Colaboratory

CRM *see CIDOC CRM*

DC Dublin Core

DCT DCMI Metadata Terms

DH Digital Humanities

DO Domain ontology

EDM Europeana Data Model

EMLO Early Modern Letters Online

FAIR Findable, Accessible, Interoperable, Reusable

FOAF Friend of a Friend

GEXF Graph Exchange XML Format

GLAM Galleries, Libraries, Archives, and Museums

GND Gemeinsame Normdatei

GKB Global Knowledge Base

GRL Genealogical Record Linkage

HISCO Historical International Standard of Classification of Occupations

HTML Hypertext Markup Language

HTTP Hypertext Transfer Protocol

HISCO Historical International Standard of Classification of Occupations

IRI Internationalized Resource Identifier

ISO International Organization for Standardization

KG Knowledge Graph

LDC Linked Data Cloud

LD Linked Data

LOD Linked Open Data

LODLAM Linked Open Data in Libraries, Archives, and Museums

NEL Named Entity Linking

NER Named Entity Recognition

NLP Natural Language Processing

PDF Portable Document Format

PNR Finnish Place Name Register (Paikannimirekisteri)

RDB Relational Database

RDF Resource Description Framework

RDFS RDF Schema

REL Relationship vocabulary

RL Record Linkage

RML RDF Mapping Language

RofL Republic of Letters

SNA Social Network Analysis

SKOS Simple Knowledge Organization System

SPARQL SPARQL Protocol and RDF Query Language

TEI Text Encoding Initiative

TF-IDF Term Frequency–Inverse Document Frequency

VIAF Virtual International Authority File

UI User Interface

ULAN Getty Union List of Artist Names

URI Uniform Resource Identifier

XML Extensible Markup Language

1. Introduction

1.1 Background and Research Environment

Semantic Web technologies and standards[1] offer solutions to solving the problems of heterogeneous, interlinked data. The underlying data model of Semantic Web is the *Resource Description Framework* (RDF) [47], in which resources are both identified and described with *Uniform Resource Identifiers* (URIs) or *Internationalized Resource Identifiers* (IRIs). This makes is possible to make a reference to an entity, e.g., a geographical location or a type of a geographical feature, by using a globally unique identifier instead of a possibly ambiguous name or a local identifier which is used only in single system. By adopting this simple principle a majority of interoperability problems in distinct geographical data sources could be avoided.

However, subsequently assigning globally unique identifiers and furthermore defining sameness for arbitrary entities found in different databases and unstructured textual materials around the world is often an extremely complicated problem. In practice materials created by humans are full of ambiguity, nuances, vague or uncertain descriptions of entities, and conflicting information. These aspect are all either hard or in some cases impossible fit into strict definitions required by data interlinking and formal logic [36]. It is clear that deep domain specific knowledge is needed when existing data sources are combined into larger wholes and re-used, often in ways which the original creators of the data never thought of. Despite these challenges, the original, somewhat idealistic Semantic Web vision of structuring "all" data into interoperable and machine-processable form offers a solid foundation for solving persistent issues of data and knowledge management, if the vision is applied in a more tightly scoped setting.

Linked Data (LD), a term coined by Tim Berners-Lee in 2006 [20, 95]

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

refers to structures and interlinked data that is made available using RDF. As RDF is a graph-based data model, RDF datasets are commonly called knowledge graphs. In the global vision of the Semantic Web [21, 22, 205] these knowledge graphs are made available on the Web in machine-understandable format and reachable by Semantic Web technologies such as the *SPARQL Protocol and RDF Query Language* (SPARQL) [253].

Biographical information, the meticulously crafted account of a person's life, holds immense value for understanding individuals, appreciating their contributions, and gaining insights into their motivations, struggles, and successes. This comprehensive information, spanning from foundational aspects like birth and death dates to personal achievements, educational background, and significant life events, forms the bedrock of understanding human experiences and the tapestry of life itself. The advent of online national biographical registers has revolutionized the access to and utilization of biographical information. These digital repositories have democratized the dissemination of biographical knowledge, making it readily available to scholars, researchers, and the general public. For instance, the *Oxford Dictionary of National Biography* (ODNB), with over 60,000 entries, stands as a testament to the transformative power of online biographical resources [161]. Similarly, the *American National Biography* (ANB) has played a pivotal role in documenting and preserving the lives of notable Americans [11]. The proliferation of online national biographical registers has enabled researchers to engage in in-depth investigations of historical figures, analyze patterns of achievement, and gain insights into the broader socio-cultural context of their lives [100]. These digital resources have also fostered cross-disciplinary collaborations, bridging the gap between traditional historical inquiry and the burgeoning field of computational social science [32, 103]. By harnessing the power of online national biographical registers, scholars can uncover new perspectives on the past, illuminate the complexities of human experiences, and inform contemporary debates on social, political, and cultural issues.

Notice that generally the term *actor* or its synonym *agent* is used when referencing to both individual people as well as to groups, such as organizations, communities, and families. However, in this thesis the main focus is on the biographical data concentrating on the individuals.

1.2 Objectives and Scope

The aim of this dissertation is summed up into the following research questions:

RQ1. How can biographical information be modeled for data analyses?

RQ2. How can biographical data be extracted, transformed, aggregated,

and enriched for a data service?

RQ3. How can biographical, prosopographical, and historiographical research be performed on the data?

RQ4. How can networks embedded in biographical data be analyzed?

Table 1.1. The relationship between the publications and research questions

Publication	RQ1	RQ2	RQ3	RQ4
I	x			
II	x			
III			x	
IV			x	
V	x	x	x	x
VI		x		x
VII			x	x
VIII	x	x	x	
IX		x		x
X	x	x		
XI		x	x	x
XII	x		x	

1.3 Research Process and Dissertation Structure

This thesis employs the *Design Science* research methodology, which diverges from traditional scientific approaches that aim to comprehend reality. Instead, design science focuses on creating artifacts that fulfill human needs and address practical problems [97, 184, 156, 85]. Rather than providing definitive solutions to the research questions, this thesis adopts a design science approach by developing tangible artifacts embedded within the projects introduced throughout the thesis. These artifacts serve as tangible expressions of the research findings and contribute to the overall understanding of the research topic.

The outcomes of design science are useful artifacts, which can be *constructs*, *models*, *methods*, or *instantiations* [97]. Within the domain of design science, constructs represent the fundamental concepts, models depict the relationships between these concepts, methods outline the procedures to achieve specific goals, and instantiations represent the concrete embodiment of an artifact in its real-world context [156]. The design science process involves identifying and defining the problem, iteratively designing, developing, testing, and evaluating the artifacts, and finally disseminating the findings to the relevant audience [184].

The artifacts examined in this thesis are the constituent parts of the developed biographical datasets. The artifacts that embody both constructs and models are the individual domain ontologies employed to represent biographical information. The biographical data models are considered model artifacts, while the individual datasets and semantic portals are instantiations. Methods are designed and employed to populate the biographical knowledge graphs. The semantic portals serve as proof-of-concept demonstrations, showcasing the suitability of the utilized artifacts for representing and utilizing biographical data as semantically harmonized data.

This thesis is structured as follows. In Chapter 2, the theoretical foundations are presented. In Chapter 3, the results in the published articles and the related data publications are reviewed and summarized. Chapter 4 delves into the broader implications of the thesis, evaluating the significance of the results, assessing the research's trustworthiness and validity, and proposing directions for future research endeavors.

2. Theoretical Foundation

The research presented in this thesis draws upon various areas and themes. This section delves into the theoretical underpinnings and related work encompassing the following domains: biographical and epistolary information, data production, data analysis, network analysis, and prosopographical research. As these research areas are wide-ranging as such, they are discussed from a Linked Open Data (LOD) centric perspective.

2.1 Biographical and Epistolary Information

A biographical dictionary is a type of reference publication that contains biographical information about notable people. Biographies form an useful source of information for the public as well as for researchers across various disciplines interested on the topic. Biographical dictionaries can be general, covering people, aka *biographees*, from all walks of life, or they might focus on people of a specific domain, such as science, politics, culture, or literature. Besides the long biographical descriptions in a free text format, the dictionaries contain short, often semi-structured biographies. Collections of individual people can also contain relations or interactions connecting the people. These connections can be in a directly announced format or they can be inferred from the data. One such a direct connection is the historical postal communication between the people in epistolary data.

2.1.1 Biographical Dictionaries and Data

Biographical dictionaries are scholarly resources used both by the public and by the academic researchers. Most dictionaries follow a traditional format which combines a lengthy non-structured text, often written with authorial individuality and personal insight, with a structured register of basic biographical facts, such as the years and places of birth and death, education, career events and achievements, products, or family relations.

If the biographical information is alike available as LOD the underlying texts can also be used for distant reading and data analysis. This data can be utilized to create intelligent user interfaces for biographical data, including tools for data visualization, analysis, and knowledge discovery in biographical, historiographical, and prosopographical research. [223].

The *Oxford Dictionary of National Biography* (ODNB) with more than 60 000 people was published in print and online in 2004 [160]. Since that pioneer work many other dictionaries have have made their editions available online. These include *The American National Biography* (ANB) [76] of the United States, *Austrian Prosopographical Information System* (APIS) [7, 19, 203], *Biography Portal* of the Netherlands, *The Dictionary of Swedish National Biography* [171], Slovenian *"Slovenska biografija"* (SBI)[1] [66, 176], and the *National Biography of Finland* (NBF) [210].

2.1.2 Biographical Databases and Knowledge Graphs

In addition to national biographies, there are international databases providing biographical data, e.g., *The Virtual International Authority File* (VIAF) [99], *Gemeinsame Normdatei* (GND) [79, 91] managed by the German National Library, and *Union List of Artist Names* (ULAN) [221, 1] by Getty Foundation. The *Biography Portal of the Netherlands* (BPN) contains biographies from 23 resources from the 18th to the 21st century and it contains 125,000 biographical descriptions of around 76,000 individuals. The project *BiographyNet* (2012–2016) uses data from the Biography Portal of the Netherlands and language technology is applied to extract entities with the connecting relations [72, 175]

For the general public there are many commercial websites concentrating on genealogical data, such as *ancestry.com*, *myheritage.com*, *Geni.com* [77, 250], and *WikiTree* [120]. In addition, Global Knowledge Bases (GKB) like *Wikidata* [248, 67] and *DBpedia* [52, 6] include actor information in addition to general information, although the level of detail of the available data may vary a lot. In addition to biographical ontologies, there are registers providing actor information, e.g., the Finnish LOD publication *KANTO* [127, 222] collected from the Finnish National Bibliography, or *HisKi* [220] collected from parish registers by the Genealogical Society of Finland.

2.1.3 Epistolary Data

In addition to the biographical data, also the communication networks formed by the interactions of the individuals can be analyzed. In a historical sense, one type of such networks can be built based on the epistolary, e.g., postal communication [164]. During the Age of Enlightenment as

[1] https://www.slovenska-biografija.si/

the postal services developed it became possible to exchange handwritten letters across Europe and beyond. This interconnectedness gave rise to the *Republic of Letters*, (Respublica Litteraria, RofL), a flourishing intellectual movement that fostered the development of modern European scientific thought, values, and institutions during the Early Modern era, spanning from the 14th to the 18th centuries [208, 63].

Data sources of early Early Modern learned correspondences are proliferating rapidly, including, e.g., The Catalogus Epistularum Neerlandicarum[2], Early Modern Letters Online (EMLO)[3] [167, 234, 236], Electronic Enlightenment[4], ePistolarium[5] [196], Europeana[6] [60, 75], the Mapping the Republic of Letters project[7], Kalliope Catalogue[8], SKILLNET[9], and correspSearch[10]. The CKCC[11] corpus stands as a Dutch counterpart to the Republic of Letters, encompassing a substantial collection of approximately 20,000 correspondences [234, 236, 108]. The *correspSearch* dataset, compiled at the Berlin-Brandenburg Academy of Sciences and Humanities, encompasses approximately 150,000 letters that have undergone scholarly editing, featuring published summaries, transcriptions, and possibly commentaries [62]. This dataset has been assembled from a vast network of content providers through the use of an XML-based Correspondence Metadata Interchange (CMI) Format, which is rooted in the Text Encoding Initiative (TEI) standards. The CMI format is essentially based on the TEI extension *correspDesc* [218].

Visualizing the epistolary data is studied in *Mapping the Republic of Letters* project[12] and in *Tudor Networks of Power*[13]. Bruneau et al. explore the application of Semantic Web Technologies to model the correspondences of French scientist *Henri Poincaré* and publish them on an online platform[14] [37]. In the case of Finland, there are collections of correspondences by renowned cultural influencers such as *Letters of Edelfelt* [212], *Elias Lönnrot Letters* [211], and *J. V. Snellman* [68].

[2] http://picarta.pica.nl/DB=3.23/

[3] http://emlo.bodleian.ox.ac.uk

[4] http://www.e-enlightenment.com

[5] http://ckcc.huygens.knaw.nl/epistolarium/

[6] http://www.europeana.eu

[7] http://republicofletters.stanford.edu

[8] http://kalliope.staatsbibliothek-berlin.de

[9] https://skillnet.nl

[10] https://correspsearch.net

[11] CKCC is an acronym for *Circulation of Knowledge: A Web-based Humanities' Collaboratory on Correspondences and Learned Practices in the 17th-century Dutch Republic*

[12] http://republicofletters.stanford.edu/

[13] http://tudornetworks.net/

[14] http://henripoincare.fr/s/correspondance/page/accueil

2.2 Linked Data

The World Wide Web offers an immense trove of biographical data, enabling researchers to employ computational methods for analysis. A good portion of this data is already organized or partially structured, making it readily accessible for systematic computerized investigation [72].

2.2.1 Publishing Data on the Web

Linked Data is a way of publishing and linking data on the internet. Unlike traditional data silos, where data is stored in separate databases and cannot be easily connected, LD uses a common vocabulary and format to allow computers to understand the relationships between different pieces of information. To foster a unified and interconnected global data landscape, Tim Berners-Lee proposed [20] four guidelines for publishing data on the Web: **1)** Utilize URIs (Uniform Resource Identifiers) as unique and enduring labels for entities, **2)** Embrace HTTP URIs, enabling individuals to readily access these identifiers, **3)** Use standards like Resource Description Framework (RDF) [47] and SPARQL (SPARQL Protocol and RDF Query Language) [57, 253] to provide comprehensive descriptions of URIs upon lookup, **4)** Leverage links associated with URIs to unveil additional pertinent information and related entities. [28]

While the traditional World Wide Web is built upon hyperlinked HTML documents, Linked Data (LD) employs a more sophisticated approach. Instead of relying solely on HTML documents, LD utilizes RDF to represent data in structured form. Unlike HTML documents that are linked by generic hyperlinks, LD employs RDF to create typed statements that connect entities in the real world, forming a network of interlinked data in a structured form. This interconnected data structure, often referred to as the *Web of Data* more accurately resembles a web of entities in the world, described by data available on the Web. [21, 28]

Data transformation from various source formats has been studied in Martinez et al. [159]. *RDB to RDF Mapping Language* (R2RML) [49] and *RDF Mapping Language* (RML) [55, 56] use customized mapping rules to transform heterogeneous data structures and serializations to RDF. The SPARQL CONSTRUCT query form[15] can be used for data mapping, transformation, or inference based on an existing RDF data publication. In programming custom transformation pipelines Python modules like *RDFLib* [197] or *Pandasrdf* [230] can be utilized. Extracting LD from textual corpora has been studied in [143, 182]. In [72] language technologies were applied to extract entities and relations in Dutch biographies for *BiographyNet*[16].

[15] https://www.w3.org/TR/rdf-sparql-query/#construct

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

2.2.2 Schemas and Ontologies for Biographical Data

Authority files [254], vocabularies, and actor ontologies serve as essential tools for identifying individuals, groups, and organizations and for representing relevant information about them. They play a pivotal role in cataloging and information management within museums, libraries, and archives. However, these tools also pose challenges for linking data across different sources due to the prevalence of alternative names, homonyms, spelling variations, diverse languages, transliteration conventions, and temporal shifts.

In the context of LD, a vocabulary serves as a standardized set of terms and definitions that governs the representation of concepts and relationships within a particular domain. It provides a common language for describing data, ensuring consistency and interoperability across different data sources. Vocabularies play a crucial role in enabling seamless data integration and analysis within the LD ecosystem. [87]

Out of the simple data schemas, e.g., *Friend-of-a-Friend* (FOAF) [33], *Relationship vocabulary* (REL) [50], *Vocabulary for Biographical Information* (BIO)[17], and *Schema.org* [89] can be used to model basic biographical information. FOAF was published in the year 2000 to model descriptions, e.g., about people and organizations where the relationships between them can be modeled using properties, such as `foaf:knows` and `foaf:member`. REL provides 35 distinct sub-properties of `foaf:knows` for modeling relations between people based, e.g., on their friendships (`rel:closeFriendOf`), professional acquaintances (`rel:collaboratesWith`), or family relationships (`rel:childOf`) [50]. *Linked Jazz* is a research project applying LOD technologies to the personal and professional lives of jazz artists. The relationship vocabulary in LinkedJazz extends REL by domain specific properties such as `linkedjazz:playedTogether` or `linkedjazz:touredWith` [182, 183, 204].

Two approaches to describe cultural heritage data are the event-centric and the object-centric schema. *The CIDOC Conceptual Reference Model* (CRM)[18] is an ISO standard for modeling cultural heritage data that has widely been used in the domains of CH, archaeology, architecture, and intangible heritage [27, 157, 81]. CRM is an event-centric framework using five upper-level type of entities: *actors*, *events*, *timespans*, *places*, and *artefacts*/*concepts*. As an example of using CRM the biographical life story of a person is modeled as a series of events taking place during a specific timespan at a known place producing an artefact [58, 158]. *Functional Requirements for Bibliographic Records* (FRBRoo) is an extension of CRM for modeling bibliographic records [59, 17].

Another approach is the object-centric schema used in the *Europeana Data Model* (EDM), where an artifact is at the center of the data model hav-

[17] `http://vocab.org/bio/`
[18] `https://cidoc-crm.org/`

ing static properties such as creator, creation date, owner and location [122]. The usage of event-centric and object-centric schemas for modeling CH data are discussed by Dijkshoorn et al. [54].

Ontology schemas have been created and implemented by biographical registers such as *Union List of Artist Names* (ULAN) [94], *Wikidata* [67, 96, 241], *Geni.com* [250], and *Gemeinsame Normdatei* (GND) [79, 91]. ULAN uses a comprehensive vocabulary of 128 terms defining inter-person and inter-actor relations. Similarly, in the data schema used by Wikidata there are properties for defining genealogical, educational, or professional connections between people. GND also uses a rich data schema with similar characteristics as CRM.

2.2.3 Semantic Disambiguation and Entity Linking

A fundamental principle of Linked Data is the ability to enhance datasets by seamlessly integrating complementary local information sources into a unified knowledge graph. This process necessitates harmonizing the data models employed by these sources and aligning the concepts and entities that populate them. *Record linkage* (RL) techniques address this latter challenge, aiming to identify matching data records across diverse databases. For instance, RL can be employed to match person records from different registries, even when those registries employ distinct metadata schemas and notational conventions, potentially representing data about the same individuals. By employing RL, richer global descriptions of persons can be constructed by merging local datasets. Furthermore, RL facilitates data enrichment by linking together local datasets that utilize different vocabularies and identifiers for representing the same resources. [39, 42, 88, 252]

Biographies of individuals can be pieced together from information gleaned from various datasets, each offering unique insights into their lives. However, the task of linking and integrating this information is challenging due to the inherent complexities of data representation. Frequently, the same full name can refer to multiple people, while different names can map to the same individual due to name changes over time. Additionally, biographical data such as birth dates may be incomplete or inaccurate, and initials may replace given names. Furthermore, the same name can be used using the Latin (as an example male name *Carolus*), Swedish (*Carl, Karl*), or Finnish (*Kaarle, Kaarlo*) variations. Furthermore, in the early 20th century it was common to change a non-Finnish family name to a new Finnish one [133].

The practices of RL are presented in [10, 88, 252]. Pattern matching using string distances is used in data linkage and duplicate detection. Statistical methods of RL are presented in [18]. Cohen et al. [45] and Christen et al. [41] provide comparizons between widely used distance

measures such as *Levenshtein* aka *Edit Distance* [148], *Jaro-Winkler* [125], or *Longest Common Substring Distance* [242]. The fuzzy string searching also overcomes some of the spelling variations and possible errors in comparing the strings. For programmers implementations of various text distance functions are available, such as *TextDistance* [179], *RapidFuzz* [8], or *Jellyfish* [229] for Python as well as deduplication packages like *Python Record Linkage Toolkit* [53] and *Splink* [149].

2.2.4 Projects Publishing Biographical Data

Genealogical research often encounters the challenge of reconciling person records from various sources. The project *AncestryAI* employs machine learning algorithms to automatically construct family trees from parish registers [152]. Antonie et al. [2] present a compelling case study of integrating *Canadian World War I* data from three distinct sources: soldier records, casualty records, and census data. Employing traditional record linkage (RL) techniques, the authors successfully merged these disparate datasets, enabling researchers to delve into the lives of those who served in the war [2]. In a similar vein, Cunningham explores the integration of military person data, combining World War I service records with census data. This integrated dataset serves as a valuable resource for analyzing the social and economic impact of the war [46]. In Ivie et al. [123] the RL process is enhanced with available genealogical data, such as information about spouses and children. Furthermore, Pixton et al. [186] utilize the pedigree information for applying a neural network for genealogical RL. Finally, constructing genealogical networks from multiple sources of vital records is discussed in several articles [64, 152, 151].

Several nationwide initiatives are underway to integrate person registries, aiming to provide comprehensive data on individuals across historical periods. The Norwegian Historical Population Register, for instance, is in the process of creating a unified registry covering the entire Norwegian population from 1800 to 1964, by merging church records and census data [224]. Antonie et al. explore the process of tracking individuals over time in 19th-century Canada for longitudinal analysis [2]. The Links project in the Netherlands seeks to reconstruct all nineteenth and early twentieth-century families by utilizing civil certificates [121, 233, 29].

In the project *Six Degrees of Francis Bacon* Early Modern social networks are inferred from biographical documents scattered across disparate books and articles. Statistical techniques and digital tools are taken up to reconstruct and visualize the historical network [244]. Elson et al. introduce [65] extracting social networks from literary fiction.

2.3 Analyzing Data and Networks

Linked Data can be viewed and analyzed by using the web-portals built upon the data or by a more detailed data analysis using programming or specific tools for data analysis. Web-based portals using faceted-search can illustrate the underlying data in the forms of Charts, Time Series, Visualizations on Maps, or Networks [26, 128, 187, 251]. LD facilitates to open new views revealing observations otherwise unachievable by traditional methods of close reading. Social networks can be constructed based on the source data with various approaches 1) using the existing relations in the data, 2) by inferring the relations based on, e.g., simultaneous events, memberships, or mutual characteristics embedded in the data, or 3) based on co-citations to people in a textual corpus such as biographical descriptions, parliamentary speeches, or in a work of literature.

2.3.1 Extracting and Analyzing Data

The underlying LD service can be accessed via its SPARQL endpoint, enabling flexible querying, analysis, and visualization of the data using various tools such as the web-based SPARQL client *Yasgui* [199], or dataset-specific query services such as *Wikidata Query Service*[19] and *Getty Vocabularies: LOD*[20]. Furthermore, by Python programming environments like *Jupyter notebooks*[21] and *Google Colaboratory* (CoLab)[22] can be used for customized data analysis where one standard workflow is to download the data in a tabular format. As an alternative for Python, *R* is an open-source statistical programming language and *RStudio* related development environment that are widely used in academia and industry for data analysis and visualization [70, 141, 247].

Christopher N. Warren states a claim that biographical dictionaries have a double voice: on one hand they describe the historical events but on the other hand they also reflect how history is studied. This claim is grounded on analysis of the *Oxford Dictionary of National Biography* (ODNB) [243]. In the article the results are represented, e.g., as timeseries, charts and tables of frequently mentioned years, the occupations of biographees parents, and mentions of people and places in biographies of different vocational groups and during the related period of time. A qualitative analysis on historical people in Wikipedia is described by Jatowt et al. [126], while Metilli et al. developed a Wikidata-based tool for constructing and visualizing narratives [166]. In [140] data analysis is performed on a cross-verified database combining biographical information from Wikipedia and Wiki-

[19] https://query.wikidata.org/

[20] https://vocab.getty.edu/queries

[21] https://jupyter.org/

[22] https://colab.research.google.com/notebooks/intro.ipynb

data. In this article also the inconsistencies between the distinct data sources are tackled as well as how much the data gets enriched when assembled using multiple datasources. In the analysis of the *Irish-language Biographical Database* (Ainm) [193], the distribution of the most common places of birth as well as vocations are analyzed [24].

2.3.2 Theoretical Background of Network Science

Network science is the study of networks, which are structures built by interconnected nodes and edges. Nodes can represent individuals, organizations, or any other type of entity, while edges represent the connections between them. The type of connections represented by an edge varies depending on the nature of the network. For example, in a social network, an edge might represent a friendship or a shared interest. Additionally, nodes can contain metadata properties, such as label, type, or location. Finally, edges can carry properties, such as weight, direction, type, or timespan. The weight of an edge might represent the strength of the relationship between the two nodes. The direction of an edge can represent the flow of information or resources between the two nodes and the type of an edge can represent the specific type of relationship between the two nodes. [174, 173]

The statistical methods of network analysis are introduced in [93, 172]. The network statistics can reveal properties on macro, meso, or micro levels. At a micro level local properties of an individual node or edge are under analysis. An *egocentric network* or ”ego network” is a type of social network that focuses on the relationships of a single individual known as ”ego”. The network includes all of the ego’s nearest neighbors aka ”alters” within a distance of a few steps, as well as the ties inside this neighborhood. By its nature, an egocentric network can be seen as an extract from a larger *sociocentric network* which focuses on all the members in the network. sociocentric networks can be used to study a wide range of social phenomena, such as the spread of information, the formation of groups, and the flow of resources. [43]

The *degree* is the most elementary measure telling about node’s connectivity and it’s adjacent neighbors. The degree indicates the number of edges connected to a node. Likewise, in a directed network, in-degree and out-degree are the number of edges directed into or out of a node mutually. *Betweenness centrality* is used to detect structural bridges and to, e.g., analyze the vulnerability of a network. *Closeness centrality* is based on the sum or average of the distances between the node and all other nodes in the graph. A node with high closeness centrality is considered central since it can quickly reach all other nodes within the network. *Eigenvector centrality*, also called *eigencentrality* or *prestige score* measures the authority, influence, or prestige of a node. *Local clustering coefficient* measures

a node's participation in a group and it is based on how well-connected its neighbors are. *PageRank* measure and its associated algorithm developed by Google were devised to organize search results according to their relevance. This approach calculates the significance of a web page iteratively, considering the frequency of external links to it and the PageRank values of the referring pages, thereby highlighting the importance of links originating from high-ranking pages. [15, 25, 34, 246, 74, 206, 245]

Typically, meso-level theories focus on analyzing populations that fall in between the individual (micro) and the collective (macro) levels. However, the term meso-level can also refer to analyses that specifically aim to bridge the gap between individual and collective phenomena. Meso-level networks typically exhibit lower connection densities and may manifest unique causal processes compared to interpersonal micro-level networks.

Finally, analyses on the macro level concentrate on global properties such as *diameter* indicating the degree of separation in the network, *connectedness* of the components, *clusters* of groups, communities, and hubs. *Global clustering coefficient* gives an overall indication of the entanglement. *Spectral analysis* is based on the eigenvalues and eigenvectors of matrices associated with the network, such as the adjacency matrix or Laplacian matrix. Spectral methods can be used for identifying the most important nodes [139] and detecting communities embedded in the networks [80, 207]. By understanding Network Analysis concepts such as *small world networks* [246], *strength of weak ties* aka *Granovetter effect* [84], *friend paradox* [69], *power-law distribution* [44], *propinquity* [9, 198], and *homophily* [162] we can gain a deeper understanding of how social networks are structured and how they function.

2.3.3 Social Network Analysis

Social Network Analysis (SNA) is a powerful research tool that delves into the intricate web of connections between interrelated entities, encompassing individuals, organizations, and entire social groups. By representing these relationships as networks, SNA offers a unique perspective for comprehending the structure and dynamics of complex social systems. Since the advent of network science around two decades ago, this field has made remarkable strides in explaining various phenomena and fundamental concepts across diverse systems, ranging from human societies to brain and cellular biology. The tools and concepts developed for network analysis enable investigations at varying levels of granularity, from examining the overall network structure to analyzing individual nodes within the network. These diagnostics, such as centrality measures, node roles, and local clustering coefficients, provide valuable insights into the network's intricacies and the roles played by its constituent elements. [246, 180, 181, 202, 239]

The concept of employing Linked Data (LD) graphs for network analysis

is a natural and intuitive extension of traditional network science methods. This approach has been explored in various research endeavors, such as the *LinkedDataLens* system, which transforms LD into a network format. The authors of this system even propose viewing LD as a "network of networks." [86]. Surendran et al. [194] utilize RDF data and SPARQL queries to perform SNA. They demonstrate how data from disparate sources can be aggregated into larger networks and enriched by each other, even through reasoning to uncover new connections between entities. SPARQL's flexibility allows for network data transformations and the creation of tabular formats suitable for further analysis. In the static approach, networks are constructed from dyadic interactions within a specified period or region. A link is established between two individuals if they have engaged in some form of contact, and the strength of the tie can be estimated based on factors such as the total number of interactions [177]. From this static perspective, researchers can investigate various network properties, including degree distribution, centrality measures, and community structures.

In Hric et al. [104] *Stochastic Block Models* are used in analyzing large-scale structures and temporal evolution of citation networks. Different genres of citation, co-citation, and bibliographic coupling networks created based on parliamentary speeches are analyzed in [189, 188]. Similarities between bibliographic coupling networks, citation networks, co-citation networks, topical networks, co-authorship networks, and co-word networks are analyzed in Yan et al. [255]. In Basu et al. relationships in the early modern theatre are analyzed by making use of databases, linguistic taggers, network analysis and visualization [14]. In Ladd et al. a network is build based on references in historical texts [136].

Many portals in the domain of Digital Humanities (DH) use network visualizations for data analysis, such as *Six Degrees of Francis Bacon*[23], *Linked Jazz*[24], and the co-citation graph and correspondent network in *ePistolarium*[25]. *LodLive*[26] allows the user to browse individual triples of a LOD database. For front-end visualizations there are libraries such as *Cytoscape.js*[27] [73], *D3.js*[28] [31], *Sigma.js*[29] [124], and *3D Force-Directed Graph*[30] [5]. *Network Navigator* [137] is an online tool for converting a CSV file into a network and to analyze the metrics. An overview of the evolution of LOD visualization is presented in [187].

Statistical SNA can be performed, e.g., using Python modules *Net-*

[23] http://sixdegreesoffrancisbacon.com
[24] https://linkedjazz.org/network/
[25] http://ckcc.huygens.knaw.nl/epistolarium
[26] http://en.lodlive.it/
[27] https://js.cytoscape.org/
[28] https://d3js.org/
[29] http://sigmajs.org/
[30] https://github.com/vasturiano/3d-force-graph

workX[31] [92], *Graph-tool*[32] [185], or the network visualization software *Gephi*[33] [13, 40]. Gephi is a versatile software platform that empowers users to explore and analyze a wide range of network data from various perspectives, including exploratory data analysis, link analysis, SNA, and biological network analysis. Furthermore, there are online tools like *Gephi Lite* [78] and *Retina* [237]. To streamline network analysis and visualization of RDF data, specialized tools are available, such as the *Semantic Web Import Plugin*[34] for Gephi. Also Web portal can facilitate SNA by providing downloading the network presentation in Gephi's GEXF format like via LinkedJazz API [204].

2.3.4 Epistolary Networks

In recent years, communication data has emerged as a valuable resource for SNA [177, 202]. The concept of representing historical epistolary data as a LOD service was first introduced in the project *Reassembling the Republic of Letters* (RofL) [228, 101, 102, 106]. Its application to digital humanities research is discussed in [108], highlighting the parallels between RofL and LOD. Various tools, data analyses, and visualizations are employed to demonstrate this analogy. In a static approach, a connection is established between two individuals if they have communicated with each other. In contrast, a temporal approach focuses on the distribution of dyadic interactions and behavioral characteristics that distinguish communication patterns.

Methods to analyze historical correspondence networks can equally well be applied to modern communication networks based on mobile or email communication or on interactions on social media platforms [202, 231]. In such cases, automated records of pairwise interactions can be transformed into communication networks, providing a valuable tool for analyzing large-scale social interactions and behavioral trends. Social signatures, which capture how individuals communicate with their alters at a given time and how this communication evolves over time, enable the assessment of the relative importance of different alters within an egocentric network. Notably, studies have revealed that social signatures tend to exhibit stability across individuals [98, 201]. In Ureña-Carrion et al. [231] patterns between contemporary mobile communication networks and epistolary networks are compared and resemblances in dyadic interactions and ego-level behaviour are found.

[31] `https://networkx.org/`
[32] `https://graph-tool.skewed.de/`
[33] `https://gephi.org/`
[34] `https://www.w3.org/2001/sw/wiki/GephiSemanticWebImportPlugin`

2.4 Prosopographical Research and Method

The term *Prosopography* was brought into a general attention by British historian Lawrence Stone (1919–1999) [209, 165]. Prosopography has its aim at understanding the dynamics of social groups, the relationships between individuals, and the social and cultural factors that shape their lives. Prosopographical research is a quantitative approach to the study of social groups, based on the collection and analysis of data on individuals.

The power of computers in processing digital data makes them ideal tools for analyzing historical figures and their connections. Their ability to identify patterns and generate statistical analyses can uncover hidden relationships and insights that would be difficult or time-consuming to detect manually. Biographical reference works, with their standardized format and consistent structure, offer a rich source of data for computational analysis [16]. These works often include individuals with shared characteristics, such as profession, gender, location, or notable achievements. This consistency allows for the identification of patterns and networks that span across different reference works and online resources [175]. Prosopographical network analysis, which utilizes this approach, has been applied to various historical periods and cultures, shedding light on the social structures of ancient societies like Rome, Greece, and China [83, 170, 200].

Prosopographical research employs a two-pronged approach: first, it identifies a group of individuals sharing specified attributes, and then, it thoroughly analyzes this group, potentially comparing it to other groups, to address the research question at hand. Faceted search on a portal serves as a tool for pinpointing the target group, while visualizations aid in scrutinizing the characteristics of the selected individuals. This iterative process enables researchers to gain a comprehensive understanding of the group's traits, patterns, and connections, shedding light on the broader historical context. [238]

3. Results

This chapter presents solutions to the research questions of the dissertation in detail. The results are compared against the current state of the art. The results as a whole are further reflected against previous research in Chapter 4. Generally, there are seven projects related to this dissertation:

- **WarSampo** (WS) is a project publishing data about the Second World War in Finland. The part of data related to this dissertation are the actor ontologies of people and the military units.
- **Norssi High School Alumni** (Norssi) publishes a register of over 10,000 alumni of the prominent Finnish high school "Norssi" between the years 1867–1992.
- **BiographySampo** (BS) provides biographical information about over 13,100 Finnish historical figures. The data is extracted from biographical texts published by the Finnish Literature Society.
- **AcademySampo** (AS) provides access to biographical data on all 28,000 Finnish and Swedish academic people educated in Finland between the years 1640–1899. The data is extracted from the student registries of the Royal Academy of Turku and the University of Helsinki.
- **LetterSampo** (LS) contains historical letter correspondences demonstrating the Dutch ePistolarium and the German CKCC data.
- **ParliamentSampo** (PS) provides a research infrastructure for studying Finnish political culture, language, and networks of Members of Parliament. The two parts of the dataset are a KG created from the minutes of plenary sessions of the Parliament of Finland [116, 217], as well as a biographical dataset representing the Finnish MPs and parliamentary organizations.
- **The Finnish Person Name Ontology** (HENKO) contains approximately 54,000 person given and family name records from the projects

BS, Norssi, and AS, as well as register data by the Finnish Digital Agency.

3.1 Modeling Biographical Information

The first research question concerns modeling a biographical information, which is the first task in creating related knowledge graphs.

How can biographical information be modeled for data analyses?

Publications I, II, V, VIII, X, and **XII** presents solutions to RQ1. by representing realized datasets of biographical information. Table 3.1 lists the published datasets, the used abbreviations, years of production, number of people, and the concerned historical time period. Table 3.2 depicts the relationships between the datasets and the published articles.

Publication I provides solutions to modeling actors, e.g., people and groups, as a part of the WS ontology. The data model follows the CIDOC CRM schema where the history of an actor is modeled using domain-specific events of military activities.

Publication II represents the data schema and the actor ontology created for the Norssi data. According to this data schema the family relationships modeled according to the Bio CRM Schema [226], and the biographical information was enriched with lifetime achievements, occupational titles, and memberships in organizations.

Publication V introduces the data model, datasets, and data service for BS where the data is compiled from five distinct datasets. According to the data schema the lifetime of a person is enriched with four types of events, and the data is rich in, e.g., including occupations, interperson references, and family relationships.

Publication X introduces the person name ontology HENKO. It is a collection of person names using the names and genders in Norssi, BS, and AS as source material. Furthermore, this data service has been utilized for gender identification in later projects such as Constellations of Correspondence (CoCo) [61, 227].

Publication XI describes the data schema for representing epistolary data in LS.

Publication XII introduces the data model of the knowledge graph (KG) used in PS. Besides the biographical data of the MPs, the publication contains a comprehensive structure of parliamentary organizations, e.g., parties, committees, and governments. Here the principles of Bio CRM were applied to model the events of memberships with different organizational roles.

Table 3.1. Overview of published datasets

Dataset	Number of People	Time Period	Published
AcademySampo (AS)	28,000	1640–1899	2021
BiographySampo (BS)	13,900	200–present	2021
LetterSampo (LS)	19,600	1510–1932	2022
Norssi High School Alumni (Norssi)	10,100	1887–1991	2017
ParliamentSampo (PS)	2800	1907–present	2023
WarSampo (WS)	100,000	1939–1945	2016

Table 3.2. The relationships between the publications and datasets

Publication	WS	Norssi	BS	AS	LS	PS
I	**x**					
II		**x**				
III		**x**	**x**			
IV	x	**x**	**x**			
V			**x**			
VI			**x**			
VII			**x**			
VIII				**x**		
IX				**x**		
X		x	x	x		
XI					**x**	
XII						**x**

3.1.1 State of the Art

Current best practices for modeling biographical data involve utilizing Linked Data and employing metadata schemas such as CIDOC CRM [58, 158] or EDM [122]. The former schema is specifically designed for harmonizing cultural heritage metadata with fine-grained event-based modeling. The latter one, on the other hand, focuses on harmonizing metadata about diverse cultural heritage collections using object-centric, event-centric, or both modeling approaches [60]. Both these models facilitate interoperability and enable seamless data harmonization. The choice of the metadata schema primarily depends on the specific research goals: whether one aims to harmonize information about historical events (CRM) or harmonize cultural heritage collection or object metadata (EDM).

In projects BS and LS, as well as in *In / Tangible European Heritage*

Table 3.3. Web Portals and Data Publications

Dataset	Data Publication	Endpoint	Portal
AcademySampo	ldf.fi/dataset/yoma	ldf.fi/yoma/sparql	akatemiasampo.fi/en
BiographySampo	ldf.fi/dataset/nbf	ldf.fi/nbf/sparql	biografiasampo.fi
LetterSampo	ldf.fi/dataset/ckcc	ldf.fi/ckcc/sparql	lettersampo.demo.seco.cs.aalto.fi
	ldf.fi/dataset/corresp	ldf.fi/corresp/sparql	
Norssi Alumni	ldf.fi/dataset/norssit	ldf.fi/norssit/sparql	www.norssit.fi/semweb
ParliamentSampo	ldf.fi/dataset/semparl	ldf.fi/semparl/sparql	parlamenttisampo.fi
WarSampo	ldf.fi/dataset/warsa	ldf.fi/warsa/sparql	www.sotasampo.fi/en

(InTaVia)[1] [129] and *ULAN* datasets from several sources were aggregated using the Provider–Proxy model, a practice originating from EDM [60, 138]. The central principle here is to aggregate the instances representing the same entity aka *Proxies* into one *Provider*. In use-cases, such as searching, or in the web portals the provider is returned or shown to the user.

The project *Golden Agents* connects many digital collections into a linked data framework resulting into a sustainable infrastructure supporting studying the relations and interactions between the producers and consumers of creative industries during the Dutch Golden Age [35]. BiographyNet contains data about over 70,000 distinct people from the Netherlands, including their birth and death dates, occupations, religions, and relationships to other people [72, 175]. The national and international LOD publications regarding parliamentary data [30, 71, 232] can contain biographical information about the MPs and ministers although the main focus as well as the public interest might be on the political speeches and documents.

3.1.2 Improving the State of the Art

Summary All the concerned projects are listed in Table 3.3. The table shows the data publication, the SPARQL service endpoint, and the online web portal for each project. The modeling schemas, data publications, and online-services bring the data following the FAIR principles available for DH research as well as for public use. Notice that the two parts of LS, CKCC and correspSearch, are two distinct data publications on separate endpoints assembled together for the data service and the web portal.

3.2 Processing Data

How can biographical data be extracted, transformed, aggregated,

[1] https://intavia.acdh-dev.oeaw.ac.at/

and enriched for a data service?

The second research question extends the first research question by focusing on the processing of the data. Publications V, VI, VIII, IX, X, and XI suggest solutions to this question by describing the methods behind the concerned datasets.

The creation of BS knowledge graph is introduced in publications V and VII. Furthermore, publication VI describes extracting and deduplicating the genealogical network of BS.

Publication VIII describes the process of converting the material of *Finnish University Student Register* (”Ylioppilasmatrikkeli” in Finnish) into the dataset AS. Publication IX describes the reconciliation of the genealogical network within that dataset. Publication X introduces using the extracted given and family names information for creating the name ontology *HENKO*. This data can be utilized for NER and NEL in automatic annotation, gender identification, as well as for enriching data in, e.g., genealogical research. Publication XII introduces creating a knowledge graph describing the Members of Parliament in Finland and related actors in politics. The RDF publication was converted from the source data in XML format provided by the Parliament of Finland.

The source data of some of the publications is aggregated from multiple sub-sources. BS is an agglomeration of five distinct data publications, AS had separate publications for the two time periods 1640–1852 and 1853–1899. LS assembled the Dutch CKCC and the German corrospSearch datasets. Finally, the actor data in WS was assembled from multiple data formats varying from OCR’d plain texts to published RDF [133].

Generally, the process pipeline in these projects has consisted of four parts. Firstly, a pattern-based information extraction from source data in a semi-structured texts. Secondly, domain-specific entities such as occupations or organizations are extracted. Thirdly, records from different parts of the source data have been disambiguated. Record linkage is introduced in publications IX and VI as well as in Koho et al. [134]. Fourthly, named entities are linked to external data publications. All publications contain linkage, e.g., to Wikidata [67, 249] and the related Wikipedia pages [12] as well as to other publications in the Sampo Series of LOD services and portals.

3.2.1 Improving the State of the Art

In general, all six introduced datasets have common features. Firstly, they consist of biographical information converted from curated source data. Secondly, they are enriched with events based on the actor-event schema. Thirdly, in all of them supporting, domain-specific ontologies such as ontologies or vocabularies of places, occupational titles, or organizations

are used for enriching the data.

The source datasets use different formats, e.g., raw text data, textual descriptions, spreadsheets, *Application Programming Interfaces* (API), *Extensible Markup Language documents* (XML), *JavaScript Object Notation* (JSON), web pages (HTML), *Portable Document Format* (PDF) documents, and RDF graphs. In many cases, the source datasets lacked standardization in terms of terminology and entity identification practices. Usually the source data was provided in two distinct formats, each requiring a tailored approach: 1) structured metadata, 2) register entries in a semi-formal format. First, the source data contained literal strings in data fields, such as given and family names, dates and places of birth and death, occupation, and related place names. Second, projects AS, BS, and Norssi provided biographical descriptions of individuals in a semi-structured format, offering additional information about family relationships and professional careers.

All the introduced data publications contain supporting ontologies in addition to the people resources. The ontologies of, e.g., places, occupations, and organizations are collected from the source data or identified from an existing ontology.

Timespans are generated using a custom-made Python module for recognizing formats of time expressions like dates, seasons, year, or decades commonly used with Finnish language. The model for the timespans adapts the four-point time schema, e.g., each timespan contains the beginning and ending times of the start and end. This model supports expressing the varying durations and the uncertainty in historical data, e.g., we only know the time of start or of the ending.

Events. In the referenced articles the events are extracted from the source material, either by extracting from biographical register descriptions (AS, BS, Norssi) or already available in the source data (LS, PS). The modeled events can be categorized into two main types, 1) basic biographical information, i.e., details of birth and death, and 2) domain specific events, e.g., lifetime events related to studies or career. Events are modeled using the superclass `crm:E5_Event` and each resource is enriched when available with links to a related Time-Span and Place.

Bio CRM is a specialized extension of CIDOC CRM designed for modeling biographical data, but it can also be applied to other Cultural Heritage data. The Bio CRM model distinguishes between enduring unary roles of actors (e.g., being a citizen or a member of a professional organization), their enduring binary relationships (e.g., being married to or having a child), and perdurant events, where participants can assume different roles from a predefined role hierarchy. The model serves as a foundation for semantic data validation and enrichment through reasoning. The enriched data conforming to Bio CRM can be effectively utilized by SPARQL queries, enabling flexible analyses that leverage the role hierarchy framework. In

the AS, BS, LS, PS, and Norssi datasets, Bio CRM schema is employed to model inter-person relationships, particularly genealogical connections. Additionally, in PS, it's utilized to represent roles held by actors in political organizations, such as member, chairman, and so forth.

Person names. Analyzing the information provided by the person names has been a crucial part of the work. This information is thereafter used for not only to detect the gender of someone or separating the given and family names but also used as a center part of disambiguation. In the datasets LS, Norssi, and WS person names are modeled using literal values with properties `familyName` or `givenName` from the vocabularies `schema` or `foaf`. Furthermore, to take the name variations into a better consideration, Simple Knowledge Organization System eXtension for Labels (SKOS-XL) [168], is used in datasets AS, BS, and PS. The properties in SKOS-XL namespace are `prefLabel`, `altLabel`, and `hiddenLabel` for preferable, alternative and hidden person names mutually. Each of the label resource contains properties for the given (`http://schema.org/givenName`) and family (`http://schema.org/familyName`) names. Handling family name additions such as nobiliary particles (*von, af, de, ...*) or genealogical prefices (*Senior, Junior*) are not implemented in the data schema or in the processing pipeline.

Referenced people. In addition to the explicit set of people, the biographees themselves the data publications contain an implicit set of people who are referenced in the data. These referenced people consist of, e.g., family members mentioned in the biographical descriptions, people mentioned in a dictionary, or biography authors. Identifying these references has been a research topic in BS as well as in AS.

Groups, e.g., organizations, families, companies, or generally all groups formed by multiple people are considered actors as well as individual people. However, biographical data is concentrated on describing the details of the individual whereas the role of groups remains secondary. In the case of Finland KANTO, and in an international context Wikidata and GND, provide comprehensive data about the organizations. In the introduced research projects the groups are extracted from the source material. In BS they were extracted from the data columns in the source spreadsheet, in AS recognized using NLP methods [216, 240] from the biographical descriptions as well as by collecting a CSV of the Student Nations. In PS the parliamentary organizations were extracted from the source data, while companies and municipal organizations were extracted or constructed based on a linkage to Wikidata entities.

Places. In most of the introduced datasets the place mentions are extracted from the source material, and furthermore linked to existing, external datasets, e.g., *Finnish Place Name Register* (PNR) [118], Wikidata, Geonames [153], TGN [94], or GND. The applied data model contains, e.g., geographical coordinate information and a spatial hierarchy.

Text Snippets and Documents. Some of the introduced datasets include textual documents like the register texts in AS and Norssi, biographical descriptions in BS, and parliamentary speeches in PS. Besides the biographical descriptions as plain text also a NLP processed graph of linguistic information is included into BS. Another example would be the contents, incipits, or summaries of a letters when modeling epistolary information.

Vocations. In most of the introduced datasets the mentions of vocations are extracted from the source material, and furthermore linked to existing, external datasets, e.g., Finnish ontology of occupations *AMMO* [130] based on *HISCO* [154, 235], Wikidata, or GND. In addition, BS included a hierarchy of occupations, and in AS the almost 10,000 occupational titles are often compounds of a related place and occupation, e.g., *Bishop of Turku* or *Merchant in Porvoo*. In PS roles of memberships in political bodies and political titles such as member, chairman, or different ministers, were also added to the ontology.

Epistolary Data In the LS project the letter correspondences are modeled using the principles outlined in the EMLO project [234]. Furthermore, each correspondence connecting a pair of actors is modeled as a LOD resource containing also addition data fields like the number of letters between a pair. In LS also the centrality measures are precalculated for the entire network of correspondences.

Shortcuts and Precalculated Values To facilitate the performance of database queries direct *shortcut* properties has been added to the RDF data. In AS the places and times of birth and death were marked as direct properties to the students in addition to the applied CIDOC CRM based model. Datasets AS, BS, and PS contain a recommender system for similar actors, these similarity values are also precalculated based on the common properties like the locations and organizations related to the lifetime events. Notice that the implemented system is stable and embedded into the data, and the web portals do not have an implementation of a recommender system based on, e.g., the browsing history of the user.

3.3 Prosopographical Data Analysis

The third research question extends the previous questions by focusing on performing prosopographical analysis both computationally as well as by making observations on the visualizations.

How can biographical, prosopographical, and historiographical research be performed on the data?

Publications III, V, VII, and XII provide results of data analysis as

answers to RQ3:

Publication III represents how prosopographical analysis can be performed in UI using faceted search and visualization tools of the BS portal.

Publication VII represents results of prosopographical analysis performed on the BS data. In this study, people were clustered, e.g., by their vocational groups and characteristics like parents' vocations, vocational groups etc.

Publication XII represents results of prosopographical analysis performed on the PS data. The results are depicted as time series and correlation matrices.

Publication IV represents examples of visualizations made with four published biographical datasets: WS, Norssi, BS, and *U.S. Congress Prosopographer* [169].

3.3.1 State of the Art

Researchers have employed linked data technologies to handle biographical information in articles [142, 175, 107]. The conference proceedings [219] encompass several papers addressing the digitization of biographical data, the application of computational methods to analyze biographies, the creation of group portraits and networks, and the development of visualizations. Implementing LOD principles in cultural heritage data management [105] and historical research [163] has emerged as a promising approach to tackle the challenges posed by isolated and semantically heterogeneous data sources. Additionally, LOD visualization is studied in several articles [26, 48, 251].

The approach of the prosopographical method is implemented by faceted search in a web portal [4, 178, 225]. First, the target group in interest is chosen. The researcher might be interested in all members of a kinship group, of an organization, neighborhood, or social class [93]. Second, the analysis are made by making conclusions and observations by the rendered tables and visualizations.

3.3.2 Improving on the State of the Art

In the publications the analysis were performed visually in the UI or by custom code implementations. The result pages in the implemented UIs serve two purposes. First, a faceted search allows the user to browse the entire set of, e.g., actors, organizations, or places. Secondly, the instance pages provide more detailed information about a chosen entity. Therefore, the instance pages could be considered as "home pages" for specified entities in the data base [114]. In the case of the faceted search, the UI allows the used to filter the target group by some common characteristics such as in a case of historical people their time of living, occupation, group, or

gender. Thereafter, the results are shown in the forms of result tables or visualized as pie charts, column charts, time series, maps, and network visualizations. The visualization tools in the UI are implementations using standard libraries providing web components like Google Charts [82], ApexCharts [3], or CytoScape.js [73].

In visualizing the data Pie and Column Charts are used for depicting the distributions of categorical variables, and Time Series for showing number of temporal, e.g., annual events. The UIs have two type of map visualizations. The first case is to show some distinct events using markers on a map and second is to show a spatial distribution of events on a heatmap. Networks are used for showing the linkage in a prosopographical group as an social network, as well as for showing a ego-centric network concentrating around a particular actor. All the utilized visualization libraries also provide interactivity, not only by the facet selection but also inside the components, e.g., clicking on a map marker can open a popup window providing extra information about the related events, or in a network rendering clicking on an alter node will change the ego of the network.

The more domain specific, custom analysis presented in the articles where made by queries in SPARQL client *Yasgui* [199] or using *Google Colaboratory*[2]. The Yasgui online application provides tools to visualize the results as charts, rendering on a map, as a gallery, or as a timeline. In order to accomplish data analysis requiring advanced processing Google Colaboratory provides all the functionalities of Python modules such as *NumPy*[3], *SciPy*[4], *NetworkX*, and *scikit-learn*[5], and furthermore, for visualization purposes, modules *matplotlib*[6] and *seaborn*[7]. The results included in the articles consist of result tables, correlation matrices, networks, and time series including, e.g., bump charts [190].

Since the data published in RDF format is already a network by its nature there is not a obvious need to render it as a network visualization showing the entities with similar characters. Instead, for many needs it remains sufficient to show the related entities as list of similar entities or recommended links. For example, in the UIs of AS and PS there are instance pages of occupations or organizations where this similarity recommendations has been implemented by querying, not only the people with the particular occupation or organization membership, but also one step further in the network, all the other occupations or organizations connected to the participated actors.

[2] `https://colab.research.google.com/notebooks/intro.ipynb`
[3] `https://numpy.org/`
[4] `https://scipy.org/`
[5] `https://scikit-learn.org/stable/`
[6] `https://matplotlib.org/`
[7] `https://seaborn.pydata.org/`

In the projects AS, BS, and PS the recommender systems were added as resources to the RDF data. The recommender system is based on precalculated similarities between the actors, in BS these values were based on the TD-IDF of the biographical description text. In AS and PS the system is estimated based the RDF data using the lifetime events [145]. In both cases the algorithm was such that it reduces the weight of most common terms and emphasizes the rarer ones. As examples of the achieved similarities and clusters it could be mentioned how in BS the spouses of the presidents of Finland were found as close matches. Likewise in AS this linkage contains a cluster of engineers who worked in Baku, Azerbaijan, for the *Oil Company Branobel* run by the Nobel brothers.

The BS publication is a collection of five distinct source datasets; 1) the National Biography of Finland, 2) Finnish Business Leaders, 3) the Finnish Generals and Admirals in the Russian armed forces 1809–1917, 4) Finnish Clergy 1554–1721, and 5) Finnish Clergy 1800–1920. As such, for the proposographical research it could be concluded that the biographical collection in BS is too heterogeneous unless a specified subgroup of people, e.g., priests, military, etc. is chosen as a target. However, by the more homogeneous nature the biographical collections of academic people in AS and members of the Parliament of Finland in PS are more suitable for prosopographical analysis.

3.4 Network Analysis

The fourth research question concerns visualizing and analyzing networks embedded in biographical data.

How can networks embedded in biographical data be analyzed?

Publications V, VI, VIII, IX, and XI provide examples as answers to this question.

Biographical data is a rich source of information about individuals' lives and relationships. This data can be used to reconstruct life histories, study social groups, and understand historical events. Network analysis is a powerful tool for analyzing biographical data. By analyzing these networks researchers can gain insights into the social, historical, and genealogical relationships between people. The networks can be analyzed visually or with computational methods to reveal the network structure with its most central nodes, components, clustering, and density.

3.4.1 Improving on the State of the Art

The networks introduced in the publications can be divided into three categories by their nature: 1) social, 2) reference, and 3) genealogical networks. The dataset LS provides a social network based on the epistolary correspondences. The datasets AS and PS provide reference networks constructed from the biographical descriptions or the parliamentary speeches. The datasets AS, BS, and PS all provide a genealogical networks of Finnish people of cultural or political significance. As a result of data processing there are two kinds of networks depending on whether they were directly extracted from the source data, or inferred from it.

Publication VI presents methods for extracting the genealogical network from the BS dataset. Publication IX presents the process of extracting a genealogical network of the AS dataset, and publication VIII results of visualizing that on a web portal. Publication VII presents network analysis performed on the reference and genealogical networks in the BS data. In this study, people were clustered by data fields, e.g., vocational groups or parents' vocations.

The genealogical networks represented in VI and IX required firstly extracting the mentioned relatives in each biographical description. Secondly, solution provided by Bio CRM schema for representing genealogical information was applied. In IX the vocabulary of genealogical relations had approx. 100 relations starting by close family relations like parent, child, or spouse, and reaching to, e.g., in-law-relatives, distant grandparents, or grandchildren many generations apart.

On-line tool Graph2SparqlServer [146] is a server-side implementation to facilitate rendering the networks in a web portal. It takes as an input two SPARQL queries, one for the links between the nodes, and the other for querying detailed metadata of the nodes. For the output there are two options: 1) a graphml formatted file or 2) a network representation in a Cytoscape [73] compatible format and in addition also calculates commonly used centrality measures, e.g., in-degree, out-degree, and PageRank. It is integrated to the Sampo-UI framework [117]. Furthermore, in the LS project also the social signatures of epistolary communication [231] are calculated in the backend server.

3.5 Results Summary

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

1. **How can biographical information be modeled for data analyses?**

Biographical data can be modeled as machine-readable Linked Open Data using Semantic Web standards. Different classes of entities are using ontologies and the entities are, if possible, linked to external databases. The presented projects are based on the actor–event schema where the basic biographical information of a person is enriched by attached events.

2. **How can biographical data be extracted, transformed, aggregated, and enriched for a data service?**

In order to transform biographical source data into LOD, a processing pipeline is used to convert the source data into rich RDF following a predefined RDF schema. The schema consists of data-specific classes of entities and the properties connecting them. A data publication can be an aggregation from several source sets, so the entities are to be disambiguated. Finally, the data is linked to external datasets not only to be enriched with additional information but most importantly in order to be a part of the global knowledge graph.

3. **How can biographical, prosopographical, and historiographical research be performed on the data?**

In the presented projects biographical, prosopographical, and historiographical research is often performed by faceted browsing on a web-portal. For example, a biographer of a historical figure can use LOD publications to learn about an individuals's family, education, career, or social relationships. Facetted browsing allows the user to filter the results by categories, such as date, place, occupation, and gender. Visualization tools are used to identify patterns and trends in the data. For more specific data or network science analysis programming tools or software can be used.

4. **How can networks embedded in biographical data be analyzed?**

Social network analysis is a powerful tool for analyzing biographical data. By analyzing networks embedded in biographical data, researchers can gain insights into the social, historical, and genealogical relationships between people. A scientist can use SNA to study the social networks of different groups of people, such as politicians, scientists, or artists. SNA can be used to identify the most influential people in each group, and to measure the level of cohesion within each group. A genealogist can use clustering to identify groups of people who are related to each other. Clique finding can be used to identify the most important families in a genealogy,

and to understand the relationships between these families.

4. Discussion

Evaluating the novel methods, tools, and implementations developed in this thesis presents a challenge due to the absence of established comparative methods. These innovative artifacts address the research problems outlined in Chapter 1. Evaluating research in the Semantic Web domain is generally complex [23], and a significant hurdle is that the effectiveness and usability of systems depend on a variety of factors, including the quality of heterogeneous source data, data handling software, and user interfaces [38].

To assess the research presented in this thesis, the following criteria have been employed: 1) theoretical implications, 2) practical implications, 3) reliability, and 4) validity. The research will be evaluated against these criteria in the following section. Finally, recommendations for future research are provided.

4.1 Theoretical Implications

The topics related to the theoretical implications were presented in the Chapter 2. Here, each research question introduced in Chapter 1 is reflected against the referenced research of the topics.

4.1.1 Modeling and Producing Biographical Data

How can biographical information be modeled for data analyses? (RQ1)

How can biographical data be extracted, transformed, aggregated, and enriched for a data service? (RQ2)

The solutions to the first and the second research questions were described in Chapter 3.1. Generally, the following four guidelines were emphasized:

1) Applying the Actor-Event schema

2) Applying the Proxy-Provider model

3) Enriching by supporting ontologies

4) Linking to external LOD cloud databases

In the following subchapters each of the four points are analyzed and discussed in detail.

The biographical collections have been modeled in RDF using mainly the CIDOC CRM model. The CIDOC CRM based actor-event schema facilitates to dealing with biographical data with sparse and heterogeneous level of provided details, e.g., in some cases a full biographical life story of an actor is available, in an another case, we might only know a few details about someone life. In addition to CIDOC CRM, shared ontologies such as DCT and FOAF are used in the schemas. The proxy-provider schema is used for assembling information gathered from distinct data sources. In this model the information about a particular actor from one data source is modeled as one of the proxies, which all are further represented by a common provider [221].

The published datasets are, in addition to the actor resources at the core, incorporated with supporting ontologies of related places, occupations, times, and relations connecting the actors. Using the supporting ontologies instead of mere literal values facilitates to, e.g., detecting prosopographical groups based on the actor characteristics, and allows implementing recommender systems with explainable connections.

The published datasets contain linkage to external LOD publications including Wikidata, Getty ULAN, GND, and VIAF as well as to other biographical publications in the Sampo series [110] and other databases in Finland. This linkage is produced not only for the actors, but as well for the supporting ontologies like places and occupations. One aim for this is creating a comprehensive biographical database of historical people in Finland.

4.1.2 Biographical Research and Network Analysis

Linked Open Data has emerged as a powerful tool for biographical, prosopographical, and historiographical research. By providing a structured and interlinked representation of historical information, LOD enables researchers to explore and analyze data in new and innovative ways.

How can biographical, prosopographical, and historiographical research be performed on the data? (RQ3)

How can networks embedded in biographical data be analyzed? (RQ4)

Biographical research involves the study of the lives of individuals and the creation of biographies. In the related articles LOD has been used to enrich biographical research by gathering, analyzing, and visualizing biographical data. Biographical information has been collected or transformed from a variety of sources, such as historical archives, biographical dictionaries, and existing online publications. This information can be used to create comprehensive biographies of individuals or to identify individuals of interest for further study. Furthermore, LOD can be used to analyze biographical patterns across different groups of individuals. This can be used to identify trends in education, occupation, social status, and other factors. Finally, LOD can be used to create visualizations of biographical data, such as charts, timeseries, networks, and maps. These visualizations can provide insights into the lives of individuals and the broader historical context in which they lived.

LOD can be used to enhance prosopographical research in several ways. Firstly, LOD can be used to create prosopographical databases that link individuals to each other and to other sources of information. This can be used to track the movement of individuals through institutions, networks, and events. Secondly, prosopographical patterns such as the distribution of individuals across different social groups, occupations, and locations can be analyzed. This can provide insights into the social structures and dynamics of a particular time period. Finally, LOD can be used to create visualizations of prosopographical data, such as social network graphs and demographic maps. These visualizations can provide a deeper understanding of the relationships between individuals and the broader society.

Network analysis of biographical data is a relatively new field, but it has the potential to make significant contributions to our understanding of history. In general, methods of SNA can be used to answer a variety of research questions about biographical and epistolary networks. For example, SNA can be used to identify influential individuals in a network, to understand how social networks form and evolve, to study the diffusion of ideas and information, and to analyze the structures in social, academic, and political networks.

4.2 Practical Implications

4.2.1 Producing Biographical Data

How can biographical information be modeled for data analyses? (RQ1)

How can biographical data be extracted, transformed, aggregated, and enriched for a data service? (RQ2)

In the case of the data conversion a semi-structured text in a register-like format has been the main source. Furthermore, a dataset as a whole can be an aggregation from multiple sub-sources, like the five subsets of BS, two of AS, and two of LS. During the conversion process the subsets have undergone harmonization since there can exist varying practices for indicating otherwise same information. Also the available information can vary by the sub-source, some may provide information about biographees' kinship relations, preferable professions, or short summary texts which might be missing from another source. Likewise, in register descriptions custom, non-standard abbreviations and symbols are often used for indicating different fields of information.

Data conversions has been multiphase processes. Custom Python scripts and modules for handling, e.g., time spans, name variations, or gender identification, have been developed, improved, and adapted based on the needs of each particular project. In addition to implementing a conversion pipeline, the data production has required pruning and wrangling with the errors and inconsistencies in the source data [150]. As an example, timespan values can be erroneous causing, e.g., negative values of duration, and the data authors might have added markings of alternative values or question marks as signs for uncertainties. To prune the data both computational and manual methods are available. Computational validation during the process [192] or RDF data validation methods [135] can be utilized. Regular expressions can be used to fix systematic errors at the character-level or word-level caused by the data acquisition, e.g., the OCR process used in the project Norssi misinterpreted Scandinavian character Å-å. Manually the data can be validated by observing outliers in, e.g., sorted data tables or in visualizations in UI. Finally, recurring undesired results of, e.g, used NER or NEL methods, can be prevented by using lists of stopwords.

4.2.2 Biographical Research and Network Analysis

How can biographical, prosopographical, and historiographical research be performed on the data? (RQ3)

How can networks embedded in biographical data be analyzed? (RQ4)

The third research question is answered by providing examples of data analysis and visualizations as shown on the web portal as well as by detailed studies performed by custom programming. The prosopographi-

cal method consists of 1) choosing the target group and 2) analyzing the results. Here, the target group is specified by the filtering done in the facet selections, and the results are shown with data tables as well as by visualizations in the forms of pie and column charts, time series, visualizations on the map, and network renderings. Furthermore, to support data analysis the web portals include landing pages for classes of entities, e.g., places, occupations, or organizations. As the results these instance pages include lists of people connected to each particular entity, e.g., of people with connections to a particular place, occupation, or organization.

BS contains a reference network built on the person references appearing in the biographical text of another person. Based on the data production the network has two kinds of links. First one of these are the HTML links manually added by the biography authors. Being so, this linkage is biased, e.g., by the order of publishing the biographies. The other type of links were computationally extracted from the biographies using methods of NLP [213]. Named entities such as mentions of people and places were extracted, and a link between the biographees were created when another person in the biography collections was mentioned in the text. Due to the slow processing of the data, the automatic annotation was not performed for the entire collection of biographies. Furthermore, the linkage was not classified, e.g., to answer why a link exists between two biographies; is it due to the biographees being colleagues, is there a teacher-student relationship, or is one the other one's role model. Altogether the reference network contains approx. 25,000 manually added HTML links and approx. 13,000 computationally inferred links.

Results of data and network analysis of BS are presented in Publication VII as well as in [215]. The results consist of tables depicting the most center actors by various centrality measures, as well as network visualizations. One observation about clustering was that the politicians and rulers tend to be clustered among themselves whereas, e.g., in the example the musicians are split into two subgroups of classical and popular music.

AS data includes a genealogical linkage by the mentioned kinship relations and a teacher-student network. The genealogical part was analyzed to detect the largest kinship clusters and longest running family lines. Likewise, the student-teacher relations were analyzed to reveal the most central figures along the timeline. Since many of the biographees were left out of these two networks and in order to perform a more profound analysis on the data a co-reference network was also build based on the similarities of the lifetime activities. In this network two students are interconnected if they share references to the same entities like organizations, places of activity, academic degrees, or professions. As an example this linkage revealed a cluster of engineers who worked in Baku, Azerbaijan for the *Oil Company Branobel* run by the Nobel brothers. Since the student register descriptions are often short, and exclude, e.g, lifetime achievements in the

cases of renowned people on their later life, the data was enriched from an external datasource choosing the Finnish Wikipedia pages. This fourth network was based on the co-citations so that a link is created when the same Wikipedia resource is referenced in the pages of two students. In most cases the linkage was based on references to a third person, place in Finland, mutual profession, or organization. An example of a common organization is the Finnish Medical Society *Duodecim* which was founded in the year 1881 by 12 Finnish physicians, all in the AcademySampo database. This result was entirely based on the information extracted from Wikipedia since *Duodecim* is not mentioned anywhere in the AS data [145]. The MP part of PS contains a similar recommender system based on the similarities of the events as AS. The recommendations are shown in the UI as table results as well as a network.

The data in LetterSampo is an aggregation of two datasets CKCC and correspSearch with their main focus on the connections built upon the mutual letter correspondences. The analysis represented in the article (Pub. XI, figures 5 and 6) shows that the datasets are different by their nature, whilst CKCC concentrates around a few central nodes, has correspSearch a larger number of hubs.

4.3 Usability, Reliability, and Validity

Individual, modular parts of the process pipeline has been evaluated using standard measures such as *precision*, *recall*, and F_1-*score* [51, 90, 155, 191]. This scientific approach is suitable for evaluating, e.g., disambiguation or linking to external databases, where a sample of results in a feasible size ($N = 50...200$) can be manually inspected. However, the usefulness and usability of the system as a whole for the general public is generally more problematic to evaluate. Out of the related projects, WarSampo has the highest amount of page views, over one million, by network statistics which is mostly explained by the public interest to the Second World War, perhaps involving their older relatives. Meanwhile, the number of page views for data publications involving people of cultural, political, or academic significance remains lower. Furthermore, BS on the second rank has over 380,000 page views, and the other publications less than 50,000 [109].

Choosing people worthwhile for a biographical collection tend to have a bias towards famous people of elite [72, 140]. This phenomena causes same renowned people to gain entries in multiple biographical collections as well as having generally the longest and most detailed biographical descriptions. Data can be incomplete due to a simple fact that only sparse pieces of information is available of someone, as well as by focusing on characteristics essential to the domain topic like the academic career or

wartime activities. In the case of Finland there is also a prevailing temporal bias towards the era of the Grand Duchy of Finland (1809–1917) and the early of 20th century. The gender distribution has a strong tendency towards male actors while women in many datasets tend to remain as mentioned family relatives of male biographees.

Figure 4.1 depicts timeseries showing the percentages of people alive for each dataset. In addition to the datasets related to this dissertation also the external datasets Wikidata and KANTO are concerned. In case of Wikidata, only people with citizenship of Finland or Grand Duchy of Finland are chosen. The values on the Y-axis are the percentages of people alive during the year specified at the X-axis. The datasets AS, BS, Norssi, and WS have their mode during the years 1850–1950. It can be noticed that in the case of WS it can be noticed that full 100 per cent of the actors were alive–due to the focus on the Second World War.

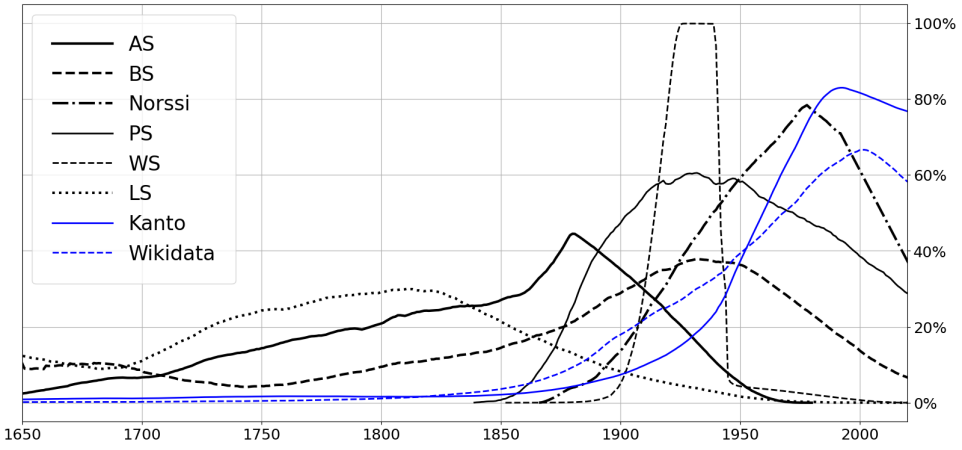


Figure 4.1. Yearly count of people alive in each data

4.4 Recommendations for Further Research

Building, interlinking, and publishing distinct biographical collections has led to constructing an interlinked, nationwide biographical data cloud concentrating on renowned Finnish historical people. The linkage connecting the data publication relating to this thesis as well as external data sources Wikidata, KANTO, and GND is depicted in Figure 4.2. The number inside each node is the corresponding number of individuals, meanwhile the numbers along the edges are the counts of connecting links. In the figure one can observe two strong triangles formed by the links, one of which is connecting Finnish datasets BS and KANTO with the resources in Wikidata. The other one is built upon the connected, international actors in LetterSampo, GND, and Wikidata.

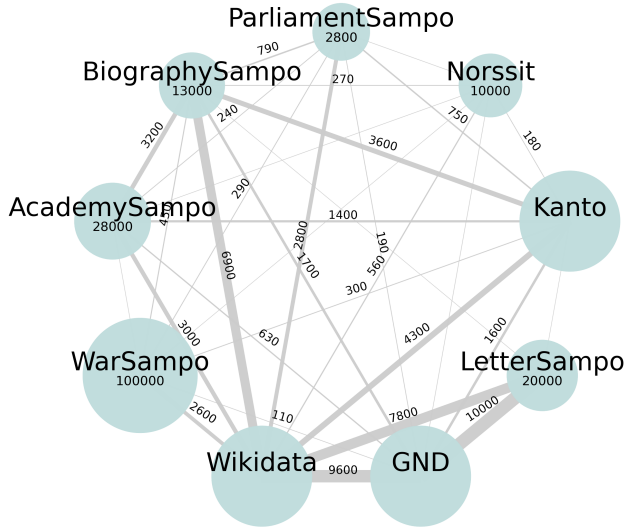


Figure 4.2. Connectivity between the actors in the central datasets

One distinct data source might be focused on a single aspect in the live of the actors such as studies, academic degrees, and careers in Norssi and AS, or activities during the wartime in WS. In addition, there can be mere mentions of, e.g., family relatives of whom more details are available in other data collections. Distinct data publications can offer information about a person's name variations, his or her lifetime whereabouts and activities, or interperson relations. Reassembling available data from previously separate data collections into one dataset would fulfill the individual life stories.

The Virtual International Authority File (VIAF) is a name authority service reassembled from 58[1] national-level collections into one web-service. One worthwhile aim for future work would be collecting the actor data in the datasets related to this thesis, as well as in other data publications in the Sampo Series, epistolary collections, and external datasets into one Finnish National Actor Ontology following the same principles of providing data as done in VIAF.

[1] 58 sources on Oct 30, 2023

Bibliography

[1] ALEXIEV, V., COBB, J., GARCIA, G., AND HARPRING, P. GVP Semantic Representation. http://vocab.getty.edu/doc//#Associative_Relationships, Accessed 8 June 2022.

[2] ANTONIE, L., GADGIL, H., GREWAL, G., AND INWOOD, K. Historical data integration a study of WWI Canadian soldiers. *ieeexplore.ieee.org*. https://ieeexplore.ieee.org/abstract/document/7836665/.

[3] APEXCHARTS. ApexCharts.js - Open Source JavaScript Charts for your website. https://apexcharts.com/, Accessed 22 November 2023.

[4] ARENAS, M., CUENCA GRAU, B., KHARLAMOV, E., MARCIUSKA, S., AND ZHELEZNYAKOV, D. Faceted search over ontology-enhanced RDF data. *dl.acm.org* (11 2014), 939–948. https://dl.acm.org/doi/abs/10.1145/2661829.2662027.

[5] ASTURIANO, V. 3d-force-graph: 3D force-directed graph component using ThreeJS/WebGL. https://github.com/vasturiano/3d-force-graph.

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

[7] AUSTRIAN CENTRE FOR DIGITAL HUMANITIES AND CULTURAL HERITAGE. Österreichisches Biographisches Lexikon. https://apis.acdh.oeaw.ac.at/, Accessed 1 September 2023.

[8] BACHMANN, M. RapidFuzz 3.6.1 documentation. https://rapidfuzz.github.io/RapidFuzz, Accessed 3 December 2023.

[9] BACKSTROM, L., SUN, E., AND MARLOW, C. Find me if you can: Improving geographical prediction with social and spatial proximity. *Proceedings of the 19th International Conference on World Wide Web, WWW '10* (2010), 61–70.

[10] BARLAUG, N., AND GULLA, J. A. Neural Networks for Entity Matching: A Survey. *ACM Transactions on Knowledge Discovery from Data (TKDD) 15*, 3 (2021), 1–37. http://dit.unitn.it/~pavel/OM/articles/BARLAUG_ACM_TKDD21.pdf.

[11] BARRETT, J. R. The American National Biography: A cornerstone of American historiography. *The American Historical Review 110*, 4 (2005), 1225–1258.

[12] BARRUS, T. MediaWiki Documentation 0.7.1. https://pymediawiki.readthedocs.io/en/latest/code.html#api, Accessed 3 June 2023.

[13] BASTIAN, M., HEYMANN, S., AND JACOMY, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. 361–362. https://www.academia.edu/download/3244556/gephi-bastian-feb09.pdf.

[14] BASU, A., HOPE, J., AND WITMORE, M. Networks and Communities in the Early Modern Theatre. *scholar.archive.org* (2017). https://scholar.archive.org/work/mv4abe7mlnhftodwgmzvrddjee/access/wayback/http://winedarksea.org/wp-content/uploads/2014/08/WH7-Networks-and-Communities.pdf.

[15] BAVELAS, A. Communication Patterns in Task-Oriented Groups. *The Journal of the Acoustical Society of America 22*, 6 (1950), 725–730.

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

[17] BEKIARI, C., DOERR, M., LE BOEUF, P., RIVA, P., AALBERG, T., BARTHÉLÉMY, J., BOUTARD, G., GÖRZ, G., IORIZZO, D., JACOB, M., LAMSFUS, C., NYMAN, M., OLIVEIRA, J., ORE, C. E., RENEAR, A. H., SMIRAGLIA, R., AND STEAD, S. Definition of FRBRoo: A conceptual model for bibliographic information in object-oriented formalism. http://repository.ifla.org/handle/123456789/659.

[18] BELL, M., AND RANADE, S. Traces through Time: a Case-study of Applying Statistical Methods to Refine Algorithms for Linking Biographical Data. *researchgate.net* (2006). https://www.researchgate.net/publication/283483529_Traces_through_time_A_case-study_of_applying_statistical_methods_to_refine_algorithms_for_linking_biographical_data.

[19] BERNÁD, Z., AND KAISER, M. The Biographical Formula: Types and Dimensions of Biographical Networks. *hcommons.org*. https://hcommons.org/deposits/item/hc:19991/.

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

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

[22] BERNSTEIN, A., HENDLER, J., AND NOY, N. A New Look at the Semantic Web. *Communications of the ACM, New York, NY, USA 59*, 9 (8 2016), 35–37. http://doi.acm.org/10.1145/2890489.

[23] 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.

[24] BHREATHNACH, U., BURKE, C., FHINN, J. M., AND CLEIRCÍN, G. A quantitative analysis of biographical data from Ainm, the Irish-language Biographical Database. http://doras.dcu.ie/23774/.

[25] BIANCHINI, M., GORI, M., AND SCARSELLI, F. Inside PageRank. *ACM Transactions on Internet Technology 5*, 1 (2 2005), 92–128.

[26] BIKAKIS, N., AND SELLIS, T. Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art. *arxiv.org* (2016). https://arxiv.org/abs/1601.08059.

[27] BINDING, C., MAY, K., AND TUDHOPE, D. Semantic interoperability in archaeological datasets: Data mapping and extraction via the CIDOC CRM. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5173 LNCS* (2008), 280–290.

[28] 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. http://eprints.epwp.eprints-hosting.org/id/eprint/92/1/bizer-heath-berners-lee-ijswis-linked-data.pdf.

[29] BLOOTHOOFT, G., CHRISTEN, P., MANDEMAKERS, K., AND SCHRAAGEN, M. Population reconstruction. *Population Reconstruction* (1 2015), 1–302.

[30] BOJĀRS, U., DARĢIS, R., LAVRINOVIČS, U., AND PAIKENS, P. LinkedSaeima: A linked open dataset of Latvia's parliamentary debates. *library.oapen.org*. https://library.oapen.org/bitstream/handle/20.500.12657/23320/1006835.pdf?sequence=1#page=66.

[31] BOSTOCK, M., OGIEVETSKY, V., AND HEER, J. D 3: Data-Driven Documents. *vis.stanford.edu*. http://vis.stanford.edu/files/2011-D3-InfoVis.pdf.

[32] BOYD, D., AND CRAWFORD, K. Critical questions for big data: Provocations for a cultural analytics community. *Information, Communication & Society 15*, 5 (2012), 662–679.

[33] BRICKLEY, D., AND MILLER, L. FOAF Vocabulary Specification 0.99. Namespace Document 14 January 2014-Paddington Edition, 2014. http://xmlns.com/foaf/spec/20140114.html.

[34] BRIN, S., AND PAGE, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer networks and ISDN systems 30*, 1-7 (1998), 107–117. https://papers.cumincad.org/data/works/att/2873.content.pdf.

[35] BROUWER, J., AND NIJBOER, H. Golden Agents. A Web of Linked Biographical Data for the Dutch Golden Age. *BD* (2017), 33–38. http://ceur-ws.org/Vol-2119/paper6.pdf.

[36] BROWN, S. Same Difference: Identity and Diversity in Linked Open Cultural Data. *euppublishing.com 16*, 1 (3 2022), 1–16. https://www.euppublishing.com/doi/abs/10.3366/ijhac.2022.0273.

[37] BRUNEAU, O., LASOLLE, N., LIEBER, J., NAUER, E., PAVLOVA, S., AND ROLLET, L. Applying and Developing Semantic Web Technologies for Exploiting a Corpus in History of Science: the Case Study of the Henri Poincaré Correspondence. *Semantic Web 12*, 2 (2021), 359–378. https://www.semantic-web-journal.net/system/files/swj2328.pdf.

[38] BURSTEIN, F., AND GREGOR, S. The Systems Development or Engineering Approach to Research in Information Systems: An Action Research Perspective. *10th Australasian Conference on Information Systems* (1999), 122–134.

[39] CHEATHAM, M., AND HITZLER, P. String similarity metrics for ontology alignment. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8219 LNCS*, PART 2 (2013), 294–309.

[40] CHERVEN, K. Network Graph Analysis and Visualization with Gephi. *kleinbauer.fr* (2013). http://kleinbauer.fr/alexis/ebook/Network_Grap_Analysis_And_Visualization_With_Gephi.pdf.

[41] CHRISTEN, P. A Comparison of Personal Name Matching: Techniques and Practical Issues. *researchgate.net* (2006), 290–294. http://dx.doi.org/10.1109/ICDMW.2006.2.

[42] CHRISTEN, P. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Science & Business Media, 2012.

[43] CHUNG, K. K., HOSSAIN, L., AND DAVIS, J. Exploring sociocentric and egocentric approaches for social network analysis. *Proceedings of the 2nd international conference on knowledge management in Asia Pacific* (2005), 1–8. https://www.academia.edu/download/40184370/Exploring_Sociocentric_and_Egocentric_Ap20151119-6475-14ez3b1.pdf.

[44] CLAUSET, A., SHALIZI, C. R., AND NEWMAN, M. E. J. Power-Law Distributions in Empirical Data. *Society for Industrial and Applied Mathematics 51*, 4 (11 2009), 661–703. https://doi.org/10.1137/070710111.

[45] COHEN, W., RAVIKUMAR, P., AND FIENBERG, S. A Comparison of String Metrics for Matching Names and Records. In *Kdd workshop on data cleaning and object consolidation* (2003), vol. 3, pp. 73–78.

[46] CUNNINGHAM, A. R. After "it's over over there": Using record linkage to enable the reconstruction of World War I veterans' demography from soldiers' experiences to civilian populations. *Historical Methods 51*, 4 (10 2018), 203–229.

[47] CYGANIAK, R., WOOD, D., LANTHALER, M., KLYNE, G., CARROLL, J. J., AND MCBRIDE, B. RDF 1.1 Concepts and Abstract Syntax. Tech. rep., World Wide Web Consortium (W3C), 2014. http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

[48] DADZIE, A.-S., AND ROWE, M. Approaches to visualising Linked Data: A survey. *Semantic Web 2*, 2 (1 2011), 89–124.

[49] DAS, S., SUNDARA, S., AND CYGANIAK, R. R2RML: RDB to RDF Mapping Language. https://www.w3.org/TR/r2rml, Accessed 16 October 2022.

[50] DAVIS, I., AND VITIELLO, E. J. RELATIONSHIP: A vocabulary for describing relationships between people. https://vocab.org/relationship/, Accessed 11 May 2023.

[51] DAVIS, J., AND GOADRICH, M. The relationship between precision-recall and ROC curves. *ACM International Conference Proceeding Series 148* (2006), 233–240.

[52] DBPEDIA ASSOCIATION. Global and Unified Access to Knowledge Graphs. https://www.dbpedia.org/, Accessed 1 September 2022.

[53] DE BRUIN, J. Python Record Linkage Toolkit. https://recordlinkage.readthedocs.io/en/latest/, Accessed 3 December 2023.

[54] DIJKSHOORN, C., AROYO, L., VAN OSSENBRUGGEN, J., AND SCHREIBER, G. Modeling cultural heritage data for online publication. *Applied Ontology 13*, 4 (11 2018), 255–271. https://doi.org/10.3233/AO-180201.

[55] DIMOU, A., AND SANDE, M. V. RDF Mapping Language (RML) Specifications. https://rml.io/specs/rml/, Accessed 16 October 2023.

[56] DIMOU, A., SANDE, M. V., SLEPICKA, J., SZEKELY, P., MANNENS, E., KNOBLOCK, C., AND VAN DE WALLE, R. Mapping Hierarchical Sources into RDF using the RML Mapping Language. *Proceedings - 2014 IEEE International Conference on Semantic Computing, ICSC 2014* (2014), 151–158. https://doi.org/10.1109/ICSC.2014.25.

[57] DODDS, L. Introducing SPARQL: Querying the Semantic Web. https://www.xml.com/pub/a/2005/11/16/introducing-sparql-querying-semantic-web-tutorial.html, Accessed 6 September 2022.

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

[59] DOERR, M., BEKIARI, C., AND LE BOEUF, P. FRBRoo, a Conceptual Model for Performing Arts. *publications.ics.forth.gr* (2008). https://publications.ics.forth.gr/_publications/drfile.2008-06-42.pdf.

[60] DOERR, M., GRADMANN, S., HENNICKE, S., ISAAC, A., MEGHINI, C., AND VAN DE SOMPEL, H. The Europeana Data Model (EDM). *World Library and Information Congress: 76th IFLA general conference and assembly Vol. 10*. https://www.academia.edu/download/37561786/149-doerretal-en.pdf.

[61] DROBAC, S., ENQVIST, J., LESKINEN, P., WAHJOE, M. F., RANTALA, H., KOHO, M., PIKKANEN, I., JAUHIAINEN, I., TUOMINEN, J., PALOPOSKI, H.-L., MELA, M. L., AND HYVÖNEN, E. The Laborious Cleaning: Acquiring and Transforming 19th-Century Epistolary Metadata. In *Digital Humanities in the Nordic and Baltic Countries Publication, DHNB2023 Conference Proceeding* (2023), vol. 5, University of Oslo Library, Norway, pp. 248–262.

[62] DUMONT, S. correspSearch -– Connecting Scholarly Editions of Letters. *Journal of the Text Encoding Initiative*, 10 (2016). https://doi.org/10.4000/jtei.1742.

[63] 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 ProjectHistorical Research in a Digital Age. *The American Historical Review 122*, 2 (4 2017), 400–424. https://academic.oup.com/ahr/article/122/2/400/3096208.

[64] EFREMOVA, J., RANJBAR-SAHRAEI, B., RAHMANI, H., OLIEHOEK, F. A., CALDERS, T., TUYLS, K., AND WEISS, G. Multi-Source Entity Resolution for Genealogical Data. *Population reconstruction* (2015), 129–154. http://www.weiss-gerhard.info/publications/MiSS_MSER_Springer_2015.pdf.

[65] ELSON, D., DAMES, N., AND MCKEOWN KATHLEEN. Extracting Social Networks from Literary Fiction. *Proceedings of the 48th annual meeting of the association for computational linguistics, ACL 2010, Uppsala, Sweden, July 11-16, 2010, no. July, Association for Computational Linguistics*, pp. 138–147. https://www.aclweb.org/anthology/P10-1015.pdf.

[66] ERJAVEC, T., DOKLER, J., AND OGRIN, P. V. Slovenian Biography. *ceur-ws.org*. https://ceur-ws.org/Vol-2119/paper3.pdf.

[67] ERXLEBEN, F., GÜNTHER, M., KRÖTZSCH, M., MENDEZ, J., AND VRANDEČIĆ, D. Introducing Wikidata to the Linked Data Web. *The Semantic Web – ISWC 2014: 13th International Semantic Web Conference* (2014), 50–65.

[68] ESKELINEN, H., HELÉN, T., KAJANDER, J., KINNUNEN, M., LINNAVALLI, S., OITTINEN, V., OJANEN, E., PITKÄRANTA, R., PULKKINEN, T., RYÖMÄ, L., SAARINEN, V., SIRONEN, E., AND TIUSANEN, A. J. V. Snellman: Kootut teokset 1–24. *Helsinki: Edita* (2001). http://snellman.kootutteokset.fi/, Accessed 22 September 2022.

[69] FELD, S. L. Why Your Friends Have More Friends than You Do1. *pdodds.w3.uvm.edu 96*, 6 (1991), 1464–77. https://pdodds.w3.uvm.edu/teaching/courses/2009-08UVM-300/docs/others/everything/feld1991a.pdf.

[70] FIELD, A., MILES, J., AND FIELD, Z. *Discovering Statistics Using R*. SAGE Publications Inc., USA, 2015.

[71] FIŠER, D., ESKEVICH, M., LENARDIČ, J., AND DE JONG, F. Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference. *Proceedings of the Workshop ParlaCLARIN III*

within the 13th Language, 2022, aclanthology.org. https://aclanthology.org/2022.parlaclarin-1.0.pdf.

[72] 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: Aspekte, Bausteine, Normen und Standards für eine europäische Biographik*, March (1 2018), 193–224. https://arxiv.org/abs/1801.07073v2.

[73] FRANZ, M., LOPES, C. T., HUCK, G., DONG, Y., SUMER, O., AND BADER, G. D. Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics 32*, 2 (1 2016), 309–311.

[74] FREEMAN, L. C. A set of measures of centrality based on betweenness. *JSTOR* (1977). https://www.jstor.org/stable/3033543.

[75] FREIRE, N., VOORBURG, R., CORNELISSEN, R., DE VALK, S., MEIJERS, E., AND ISAAC, A. Aggregation of Linked Data in the Cultural Heritage Domain: A Case Study in the Europeana Network. *openreview.net*. https://openreview.net/pdf?id=4P9jWXGAfZ3.

[76] GARRATY, J. A., AND CARNES, M. C. American national biography. *Oxford University Press* (1999). https://global.oup.com/academic/product/american-national-biography-9780195206357.

[77] GENI.COM. Family Tree & Family History at Geni.com. https://www.geni.com/, Accessed 1 September 2022.

[78] GEPHI. Gephi Lite. https://gephi.org/gephi-lite/, Accessed 14 November 2023.

[79] GERMAN NATIONAL LIBRARY. DNB - The Integrated Authority File (GND). https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html, Accessed 1 September 2019.

[80] GIRVAN, M., AND NEWMAN, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America 99*, 12 (6 2002), 7821–7826. https://www.pnas.org/doi/abs/10.1073/pnas.122653799.

[81] GOFF, E. L., MARLET, O., AND RODIER, X. Interoperability of the ArSol (Archives du Sol) database based on the CIDOC-CRM ontology. *CAA2014: 21st Century Archaeology: Concepts, methods and tools, 2015, torrossa.com*. https://www.torrossa.com/gs/resourceProxy?an=5245172&publisher=FZR707#page=191.

[82] GOOGLE FOR DEVELOPERS. Chart Gallery, Charts, Google for Developers. https://developers.google.com/chart/interactive/docs/gallery.

[83] GRAHAM, S., AND RUFFINI, G. Network Analysis and Greco-Roman Prosopography. *hcommons.org*. https://hcommons.org/deposits/view/hc:18912/CONTENT/network_analysis_and_greco-roman_prosopo.pdf/.

[84] GRANOVETTER, M. S. The Strength of Weak Ties. *American journal of sociology 78*, 6 (1973), 1360–1380. https://info.sice.indiana.edu/~katy/L597-F05/granovetter73.pdf.

[85] GREGOR, S., AND HEVNER, A. R. Positioning and Presenting Design Science Research for Maximum Impact. *MIS quarterly 37*, 2 (6 2013), 337–355.

[86] GROTH, P., AND GIL, Y. Linked Data for Network Science. *Citeseer*. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.224.1421&rep=rep1&type=pdf.

[87] GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition 5*, 2 (6 1993), 199–220.

[88] GU, L., BAXTER, R., VICKERS, D., AND RAINSFORD, C. Record linkage: Current practice and future directions. *Citeseer*. https://www.researchgate.net/profile/Lifang-Gu/publication/2479819_Record_Linkage_Current_Practice_and_Future_Directions/links/0deec52969a26cf1b0000000/Record-Linkage-Current-Practice-and-Future-Directions.pdf.

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

[90] HACHEY, B., RADFORD, W., NOTHMAN, J., HONNIBAL, M., AND CURRAN, J. R. Evaluating entity linking with Wikipedia. *Artificial Intelligence 194* (jan 2013), 130–150. http://dx.doi.org/10.1016/j.artint.2012.04.005.

[91] HAFFNER, A. GND Ontology. https://d-nb.info/standards/elementset/gnd, Accessed 16 December 2022.

[92] HAGBERG, A., SWART, P. J., AND SCHULT, D. A. Exploring network structure, dynamics, and function using NetworkX. https://www.osti.gov/servlets/purl/960616.

[93] HANNEMAN, R. A., AND RIDDLE, M. Introduction to social network methods. http://wiki.gonzaga.edu/dpls707/images/6/6e/Introduction_to_Social_Network_Methods.pdf, Accessed 7 October 2021.

[94] HARPRING, P. Development of the Getty Vocabularies: AAT, TGN, ULAN, and CONA. *The University of Chicago Press Journals 29*, 1 (9 2015), 67–72. https://doi.org/10.1086/adx.29.1.27949541.

[95] HEATH, T., AND BIZER, C. Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology 1*, 1 (2011), 1–121.

[96] HERNÁNDEZ, D., HOGAN, A., AND KRÖTZSCH, M. Reifying RDF: What Works Well with Wikidata? *ceur-ws.org* (2014). http://ceur-ws.org/Vol-1457/SSWS2015_paper3.pdf.

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

[98] HEYDARI, S., ROBERTS, S. G., DUNBAR, R. I. M., AND SARAMÄKI, J. Multichannel social signatures and persistent features of ego networks. *Applied Network Science 3*, 1 (5 2018). https://doi.org/10.1007/s41109-018-0065-4.

[99] HICKEY, T. B., AND TOVES, J. A. Managing ambiguity in VIAF. *mirror.dlib.org*. http://mirror.dlib.org/dlib/july14/hickey/07hickey.print.html.

[100] HOGAN, T. D., MATTHEWS, J., SWARTZ, N., AND ZIMMER, M. Patterns of achievement in online biographical reference works. *The Journal of American History 103*, 1 (2016), 162–182.

[101] HOTSON, H., AND HYVÖNEN, E. Topics. In *Reassembling the Republic of Letters in the Digital Age*, H. Hotson and T. Wallnig, Eds. Göttingen University Press, 2019, ch. 2.5, pp. 137–148. https://doi.org/10.17875/gup2019-1146.

[102] HOTSON, H., WALLNIG, T., TUOMINEN, J., EETU MÄKELÄ, AND HYVÖNEN, E. People. In *Reassembling the Republic of Letters in the Digital Age*, H. Hotson and T. Wallnig, Eds. Göttingen University Press, 2019, ch. 2.4, pp. 119–136. https://doi.org/10.17875/gup2019-1146.

[103] HOWARD, P. N. Computational social science and the political imagination. *Computational Social Science 5*, 2 (2018), 186–192.

[104] HRIC, D., KASKI, K., AND KIVELÄ, M. Stochastic block model reveals maps of citation patterns and their evolution in time. *Journal of Informetrics 12*, 3 (8 2018), 757–783.

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

[106] HYVÖNEN, E., AHNERT, R., AHNERT, S. E., TUOMINEN, J., MÄKELÄ, E., LEWIS, M., AND FILARSKI, G. Reconciling metadata. In *Reassembling the Republic of Letters in the Digital Age*, H. Hotson and T. Wallnig, Eds. Göttingen University Press, 2019, ch. 3.2, pp. 223–235. https://doi.org/10.17875/gup2019-1146.

[107] HYVÖNEN, E., ALONEN, M., AND IKKALA, E. Life Stories as Event-based Linked Data: Case Semantic National Biography. *seco.cs.aalto.fi*. http://seco.cs.aalto.fi/publications/2014/hyvonen-et-al-life-stories-iswc-2014.pdf.

[108] HYVÖNEN, E., LESKINEN, P., AND TUOMINEN, J. LetterSampo–Historical Letters on the Semantic Web: A Framework and Its Application to Publishing and Using Epistolary Data. https://seco.cs.aalto.fi/publications/2020/hyvonen-et-al-lettersampo-2020.pdf.

[109] HYVÖNEN, E. Creating and Using Biographical Dictionaries for Digital Humanities Based on Linked Data: A Survey of Web Services in Use in Finland. In *Biographical Data in a Digital World 2022 (BD 2022), Tokyo* (August 2023), Proceedings, accepted.

[110] HYVÖNEN, E. Digital Humanities on the Semantic Web: Sampo Model and Portal Series. *Semantic Web – Interoperability, Usability, Applicability 14*, 4 (2023), 729–744. https://doi.org/10.3233/SW-223034.

[111] HYVÖNEN, E., ALONEN, M., IKKALA, E., AND MÄKELÄ, E. Life Stories as Event-based Linked Data: Case Semantic National Biography. In *Proceedings of ISWC 2014 Posters & Demonstrations Track* (October 2014), CEUR Workshop Proceedings. http://ceur-ws.org/Vol-1272/.

[112] HYVÖNEN, E., LESKINEN, P., HEINO, E., TUOMINEN, J., AND SIROLA, L. Reassembling and Enriching the Life Stories in Printed Biographical Registers: Norssi High School Alumni on the Semantic Web. In *Proceedings, Language, Data and Knowledge (LDK 2017)* (June 2017), Springer-Verlag, pp. 113–119. https://link.springer.com/chapter/10.1007/978-3-319-59888-8_9.

[113] HYVÖNEN, E., LESKINEN, P., TAMPER, M., RANTALA, H., IKKALA, E., TUOMINEN, J., AND KERAVUORI, K. BiographySampo - Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research. In *The Semantic Web. ESWC 2019* (June 2019), P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A. J. Gray, V. Lopez, A. Haller, and K. Hammar, Eds., Springer-Verlag, pp. 574–589. https://doi.org/10.1007/978-3-030-21348-0_37.

[114] 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 *The Fourth Digital Humanities in the Nordic Countries 2019 (DHN2019), Book of Abstracts* (March 2019), University of Copenhagen. https://cst.dk/DHN2019Pro/abstracts/hyvonen-et-al-dhn-2019-bs.pdf.

[115] 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)* (2022), CEUR-WS Proceedings, vol. 3152, pp. 16–23. https://ceur-ws.org/Vol-3152/BD2019_paper_3.pdf.

[116] HYVÖNEN, E., SINIKALLIO, L., LESKINEN, P., MELA, M. L., TUOMINEN, J., ELO, K., DROBAC, S., KOHO, M., IKKALA, E., TAMPER, M., LEAL, R., AND KESÄNIEMI, J. Linked Data Approach for Studying Parliamentary Speeches and Networks of Politicians in Finland 1907-2021 (long paper). In *Digital Humanities 2022, Conference Abstracts, July 25-29, 2022 Online, Tokyo. Japan, University of Tokyo* (July 2022), ADHO, pp. 254–257. https://dh2022.adho.org/.

[117] IKKALA, E., HYVÖNEN, E., RANTALA, H., AND KOHO, M. Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. *Semantic Web 13*, 1 (1 2022), 69–84. http://rdf.js.org.

[118] IKKALA, E., HYVÖNEN, E., AND TUOMINEN, J. An Ontology of World War II Places for Linking and Enriching Heterogeneous Historical Data Sources. In *17th International Conference of Historical Geographers (ICHG 2018), Book of Abstracts* (Warsaw, Poland, 7 2018), no. 194.

[119] IKKALA, E., KOHO, M., HEINO, E., LESKINEN, P., HYVÖNEN, E., AND AHORANTA, T. Prosopographical Views to Finnish WW2 Casualties Through Cemeteries and Linked Open Data. In *Proceedings of the Workshop on Humanities in the Semantic Web (WHiSe II)* (October 2017), CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2014/.

[120] INTERESTING.COM, INC. WikiTree: The Free Family Tree. https://www.wikitree.com/, Accessed 1 September 2023.

[121] INTERNATIONAL INSTITUTE OF SOCIAL HISTORY. LINKing System for historical family reconstruction (LINKS). https://iisg.amsterdam/en/hsn/projects/links, Accessed 4 October 2023.

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

[123] IVIE, S., HENRY, G., GATRELL, H., AND GIRAUD-CARRIER, C. A metric-based machine learning approach to genealogical record linkage. *Citeseer*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.182.6051&rep=rep1&type=pdf.

[124] JACOMY, A. jacomyal/sigma.js: A JavaScript library aimed at visualizing graphs of thousands of nodes and edges. https://github.com/jacomyal/sigma.js.

[125] JARO, M. A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association 84* (1989), 414–420. https://www.tandfonline.com/action/journalInformation?journalCode=uasa20.

[126] 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 (12 2019), 287–305. https://link.springer.com/article/10.1007/s00799-018-0231-4.

[127] KANSALLISKIRJASTO. Finto: KANTO - Kansalliset toimijatiedot. https://finto.fi/finaf/en/?clang=fi, Accessed 21 December 2021.

[128] KEHRER, J., AND, H. H. I. T. O. V., AND 2012, U. Visualization and visual analysis of multifaceted scientific data: A survey. *ieeexplore.ieee.org 19*, 3 (2013). https://ieeexplore.ieee.org/abstract/document/6185547/.

[129] KESÄNIEMI, J., SCHLÖGL, M., TUOMINEN, J., DE BOER, V., AND SUGIMOTO, G. Towards Reusable Aggregated Biographical Research Data: Provenance and Versioning in the InTaVia Knowledge Graph. In *Digital Humanities in the Nordic and Baltic Countries Seventh Conference (DHNB 2023), Book of Abstracts* (March 2023), S. Gilbert and A. Rockenberger, Eds., University of Oslo Library, Oslo, Norway, p. 117. https://doi.org/10.5281/zenodo.7670464.

[130] KOHO, M., GASBARRA, L., TUOMINEN, J., RANTALA, H., JOKIPII, I., AND HYVÖNEN, E. AMMO Ontology of Finnish Historical Occupations. In *Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH'19)* (June 2019), A. Poggi, Ed., vol. 2375, CEUR Workshop Proceedings, pp. 91–96. http://ceur-ws.org/Vol-2375/.

[131] KOHO, M., AND HYVÖNEN, E. Studying Occupations and Social Measures of Perished Soldiers in WarSampo Linked Open Data. In *Biographical Data in a Digital World 2022 (BD 2022), Tokyo* (August 2023), CEUR Workshop Sproceedings. Forth-coming.

[132] KOHO, M., IKKALA, E., AND HYVÖNEN, E. Reassembling the Lives of Finnish Prisoners of the Second World War on the Semantic Web. In *Proceedings of the Third Conference on Biographical Data in a Digital World (BD 2019)* (June 2022), CEUR Workshop Proceedings, pp. 31–39. http://ceur-ws.org/Vol-3152/BD2019_paper_5.pdf.

[133] KOHO, M., IKKALA, E., LESKINEN, P., TAMPER, M., TUOMINEN, J., AND HYVÖNEN, E. WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data. *Semantic Web – Interoperability, Usability, Applicability 12*, 2 (1 2021), 265–278. https://doi.org/10.3233/SW-200392.

[134] KOHO, M., LESKINEN, P., AND HYVÖNEN, E. Integrating Historical Person Registers as Linked Open Data in the WarSampo Knowledge Graph. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2020), vol. 12378 LNCS.

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

[136] LADD, J. R., HOUSTON, N., AND ECKERT, L. Imaginative Networks: Tracing Connections Among Early Modern Book Dedications. *scholar.archive.org* (2021). https://scholar.archive.org/work/nsbyjoiok5helasjqmvxtr7674/access/wayback/https://culturalanalytics.org/article/21993.pdf.

[137] LADD, J. R., AND LEBLANC, Z. Network Navigator. https://networknavigator.jrladd.com/.

[138] LAGOZE, C., DE SOMPEL, H., NELSON, M. L., WARNER, S., SANDERSON, R., AND JOHNSTON, P. Object re-use & exchange: A resource-centric approach. *arXiv preprint arXiv:0804.2273* (2008).

[139] LANGVILLE, A. N., AND MEYER, C. D. A survey of eigenvector methods for web information retrieval. *SIAMAN Langville, CD MeyerSIAM review, 2005, SIAM 47*, 1 (3 2005), 135–161. https://epubs.siam.org/doi/abs/10.1137/S0036144503424786.

[140] LAOUENAN, M., BHARGAVA, P., EYMÉOUD, J.-B., GERGAUD, O., PLIQUE, G., AND WASMER, E. A cross-verified database of notable people, 3500BC-2018AD. *Scientific Data 9*, 1 (2022), 290. https://doi.org/10.1038/s41597-022-01369-4.

[141] LAROSE, C. D., AND LAROSE, D. T. Data science using Python and R. https://toc.library.ethz.ch/objects/pdf03/z01_978-1-119-52681-0_01.pdf.

[142] LARSON, R. R. Demonstration: Bringing Lives to Light: Browsing and Searching Biographical Information with a Metadata Infrastructure. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 4675 LNCS* (2007), 539–542. https://link.springer.com/chapter/10.1007/978-3-540-74851-9_63.

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

[144] LESKINEN, P., AND HYVÖNEN, E. Using the AcademySampo Portal and Data Service for Biographical and Prosopographical Research in Digital Humanities. In *ISWC-Posters-Demos-Industry 2021 International Semantic Web Conference (ISWC) 2021: Posters, Demos, and Industry Tracks* (Oct 2021), CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2980/paper330.pdf.

[145] LESKINEN, P., AND HYVÖNEN, E. Biographical and Prosopographical Analyses of Finnish Academic People 1640–1899 Based on Linked Open Data. In *Proceedings of the Biographical Data in a Digital World 2022 (BD 2022)* (August 2023), ZRC SAZU, Založba ZRC. https://doi.org/10.3986/9789610508120_7.

[146] LESKINEN, P., HYVÖNEN, E., AND TUOMINEN, J. Sparql2GraphServer: a Server-side Tool for Extracting Networks from Linked Data for Data Analysis. In *ISWC-Posters-Demos-Industry 2021 International Semantic Web Conference (ISWC) 2021: Posters, Demos, and Industry Tracks* (Oct 2021), CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2980/paper343.pdf.

[147] LESKINEN, P., RANTALA, H., AND HYVÖNEN, E. Analyzing the Lives of Finnish Academic People 1640–1899 in Nordic and Baltic Countries: AcademySampo Data Service and Portal. In *DHNB 2022 The 6th Digital Humanities in Nordic and Baltic Countries Conference* (March 2022), CEUR Workshop Proceedings, long papers, Vol. 3232. http://ceur-ws.org/Vol-3232/paper07.pdf.

[148] LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (1966), vol. 10, Soviet Union, pp. 707–710. https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf.

[149] LINACRE, R., LINDSAY, S., MANASSIS, T., SLADE, Z., HEPWORTH, T., KENNEDY, R., AND BOND, A. Splink: Free software for probabilistic record linkage at scale. *International Journal of Population Data Science 7*, 3 (Aug. 2022).

[150] MÄKELÄ, E., LAGUS, K., LAHTI, L., SÄILY, T., TOLONEN, M., HÄMÄLÄINEN, M., KAISLANIEMI, S., AND NEVALAINEN, T. Wrangling with Non-Standard Data. https://www.academia.edu/download/63601258/wrangling_non_standard_data20200611-122546-1nzwdwa.pdf.

[151] MALMI, E. Collective Entity Resolution Methods for Network Inference. https://aaltodoc.aalto.fi/handle/123456789/31841.

[152] MALMI, E., RASA, M., AND GIONIS, A. Ancestryai: A tool for exploring computationally inferred family trees. *26th International World Wide Web Conference 2017, WWW 2017 Companion* (2017), 257–261. http://dx.doi.org/10.1145/3041021.3054728.

[153] MALTESE, V., AND FARAZI, F. A semantic schema for GeoNames. http://eprints.biblio.unitn.it/4088/.

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

[155] MANNING, C. D., RAGHAVAN, P., AND SCHÜTZE, H. Introduction to Information Retrieval, Cambridge University Press. 2008. ISBN-13 978-0-521-86571-5. *cambridge.org*. https://doi.org/10.1017/S1351324909005129.

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

[157] MARLET, S., CURET, X., AND RODIER, B. B.-M. Using CIDOC CRM for Dynamically Querying ArSol, a Relational Database, from the Semantic Web. *Proceedings of the 43rd annual conference on computer applications, books.google.com* (2016).

[158] MARTIN DOERR. The CIDOC CRM, an ontological approach to schema heterogeneity. *drops.dagstuhl.de* (2005). https://drops.dagstuhl.de/opus/volltexte/2005/35/.

[159] MARTINEZ-RODRIGUEZ, J. L., HOGAN, A., AND LOPEZ-AREVALO, I. Information Extraction meets the Semantic Web: A Survey. *semantic-web-journal.net* (2016). http://semantic-web-journal.net/system/files/swj1744.pdf.

[160] MATTHEW, H. C. G., HARRISON, B., AND LONG, R. J. The Oxford dictionary of national biography. https://digitalcommons.fairfield.edu/philosophy-books/20/.

[161] MATTHEWS, J. The transformative power of online biographical resources. *The History Teacher 44*, 4 (2011), 473–491.

[162] MCPHERSON, M., SMITH-LOVIN, L., AND COOK, J. M. Birds of a Feather: Homophily in Social Networks. *Annual review of sociology 27*, 1 (2001), 415–444. https://www.researchgate.net/publication/200110353.

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

[164] MERRIAM-WEBSTER. Definition & Meaning: Epistolary. https://www.merriam-webster.com/dictionary/epistolary, Accessed 14 December 2022.

[165] MERRIAM-WEBSTER. Definition & Meaning: Prosopography. https://www.merriam-webster.com/dictionary/prosopography, Accessed 14 December 2022.

[166] METILLI, D., BARTALESI, V., AND MEGHINI, C. A Wikidata-based tool for building and visualising narratives. *International Journal on Digital Libraries* (1 2019). https://doi.org/10.1007/s00799-019-00266-3.

[167] MIERT, D. V. What was the Republic of Letters? *ugp.rug.nl*. https://ugp.rug.nl/groniek/article/download/27601/25014/.

[168] MILES, A., AND BECHHOFER, S. SKOS eXtension for Labels (SKOS-XL) Namespace Document - HTML Variant, 18 August 2009 Recommendation Edition. https://www.w3.org/TR/skos-reference/skos-xl.html, Accessed 17 December 2021.

[169] MIYAKITA, G., LESKINEN, P., AND HYVÖNEN, E. U.S. Congress prosopographer - A tool for prosopographical research of legislators. In *CEUR Workshop Proceedings* (2018), vol. 2180.

[170] NAILS, D. The People of Plato: A Prosopography of Plato and Other Socratics. https://philpapers.org/rec/NAITPO-4https://hackettpublishing.com/pdfs/The_People_of_Plato_Errata_and_Addenda_2.pdf.

[171] NATIONAL ARCHIVES OF SWEDEN. Svenskt Biografiskt Lexikon. https://sok.riksarkivet.se/Sbl/Start.aspx?lang=en, Accessed 1 September 2022.

[172] NETTLETON, D. F. Data mining of social networks represented as graphs. *Computer Science Review 7*, 1 (2 2013), 1–34.

[173] NEWMAN, M. E. J. The structure and function of complex networks. *cs.rice.edu*. https://www.cs.rice.edu/~nakhleh/COMP572/Material/StructureAndFunctionOfComplexNetworks.pdf.

[174] NEWMAN, M. E. J. Networks: An Introduction. *Oxford University Press* (9 2010), 1–784. https://doi.org/10.1093/oso/9780198805090.001.0001.

[175] 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. *Citeseer*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.429.1506&rep=rep1&type=pdf.

[176] OGRIN, P. V. Slovenian Biographical Lexicon - From a Digital Edition to an On-Line Application. https://www.academia.edu/8324784/Slovenian_Biographical_Lexicon_From_a_Digital_Edition_to_an_On_Line_Application.

[177] ONNELA, J. P., SARAMÄKI, J., HYVÖNEN, J., SZABÓ, G., LAZER, D., KASKI, K., KERTÉSZ, J., AND BARABÁSI, A. L. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences of the United States of America 104*, 18 (5 2007), 7332–7336.

[178] OREN, E., DELBRU, R., AND DECKER, S. Extending Faceted Navigation for RDF data. In *The Semantic Web – ISWC 2006: 5th International Semantic Web Conference* (Athens, GA, USA, 11 2006), I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. M. Aroyo, Eds., vol. 4273 of *Lecture Notes in Computer Science*, pp. 559–572.

[179] ORSINIUM. Textdistance 4.6.0. https://pypi.org/project/textdistance/, Accessed 10 November 2022.

[180] 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.

[181] PANDA, M., EL-BENDARY, N., SALAMA, M. A., HASSANIEN, A. E., AND ABRAHAM, A. Computational social networks: Tools, perspectives, and challenges. *Computational Social Networks: Tools, Perspectives and Applications 9781447140481* (8 2012), 3–23.

[182] PATTUELLI, C. Linked Jazz 52nd Street: A LOD Crowdsourcing Tool to Reveal Connections among Jazz Artists, 2013. https://www.academia.edu/4091708/Linked_Jazz_52nd_Street_A_LOD_Crowdsourcing_Tool_to_Reveal_Connections_among_Jazz_Artists.

[183] PATTUELLI, M. C., WELLER, C., AND SZABLYA, G. Linked Jazz: an exploratory pilot. *dcpapers.dublincore.org* (2011). http://dcpapers.dublincore.org/pubs/article/view/3637.

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

[185] PEIXOTO, T. P. The Graph-tool Python library. *Figshare* (2014). DOI 10.6084/m9.figshare.1164194.

[186] PIXTON, B., AND GIRAUD-CARRIER, C. Using structured neural networks for record linkage. *Proceedings of the sixth annual workshop on technology for family history and genealogical research.* (2006). https://www.researchgate.net/publication/267718679_Using_Structured_Neural_Networks_for_Record_Linkage.

[187] PO, L., BIKAKIS, N., DESIMONI, F., AND PAPASTEFANATOS, G. Linked Data Visualization Techniques, Tools, and Big Data. *Synthesis Lectures on the Semantic Web: Theory and Technology 10*, 1 (3 2020), 1–157.

[188] POIKKIMÄKI, H., LESKINEN, P., AND HYVÖNEN, E. Applying Network and Bibliometric Analyses to Mentions of Politicians in Plenary Speeches: Case ParliamentSampo - Parliament of Finland on the Semantic Web.

[189] POIKKIMÄKI, H., LESKINEN, P., TAMPER, M., AND HYVÖNEN, E. Analyses of networks of politicians based on linked data: Case parliamentsampo - parliament of finland on the semantic web. In *New Trends in Database and Information Systems* (August 2022), Springer International Publishing, pp. 585–592.

[190] PONNE, B. Towards Data Science, Create Bump Charts With Matplotlib. https://towardsdatascience.com/create-bump-charts-with-matplotlib-431b0e6fcb90, Accessed 22 November 2023.

[191] POWERS, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* (10 2020). http://arxiv.org/abs/2010.16061.

[192] PYDANTIC. Welcome to Pydantic. https://docs.pydantic.dev/latest/, Accessed 29 November 2023.

[193] RAGHALLAIGH, B. O., AND CLEIRCÍN, G. Ó. Ainm.ie: Breathing new life into a canonical collection of irish-language biographies. In *BD* (2015), pp. 20–23. http://ceur-ws.org/Vol-1399/paper4.pdf.

[194] RAJI, P. S., AND SURENDRAN, S. RDF approach on social network analysis. *International Conference on Research Advances in Integrated Navigation Systems, RAINS 2016* (12 2016).

[195] RANTALA, H., AND HYVÖNEN, E. Who is Related to What and How? Using Biographical Knowledge Graphs for Explainable Relational Search in BiographySampo. Submitted for review.

[196] RAVENEK, W., HEUVEL, C. V. D., AND GERRITSEN, G. The ePistolarium: Origins and Techniques. *JSTOR*. https://www.jstor.org/stable/j.ctv3t5qjk.33.

[197] RDFLIB TEAM. RDFLib 7.0.0 Documentation. https://rdflib.readthedocs.io/en/stable/, Accessed 10 October 2023.

[198] REAGANS, R. Close encounters: Analyzing how social similarity and propinquity contribute to strong network connections. *pubsonline.informs.org 22*, 4 (2011), 835–849. https://pubsonline.informs.org/doi/abs/10.1287/orsc.1100.0587.

[199] RIETVELD, L., AND HOEKSTRA, R. The YASGUI Family of SPARQL clients. *Semantic Web – Interoperability, Usability, Applicability 8*, 3 (2017), 373–383. https://www.semantic-web-journal.net/system/files/swj1126.pdf.

[200] RUDOLPH, H., AND CHEN, S. Biography for Historical Analysis: A Chinese Biographical Database. *Journal of Historical Network Research 5*, 1 (9 2021). http://jhnr.uni.lu/index.php/jhnr/article/view/124.

[201] SARAMÄKI, J., LEICHT, E. A., LOPEZ, E., ROBERTS, S. G. B., REED-TSOCHAS, F., AND DUNBAR, R. I. M. Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences 111*, 3 (1 2014), 942–947. https://doi.org/10.1073/pnas.1308540110.

[202] SARAMÄKI, J., AND MORO, E. From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. *The European Physical Journal B 2015 88:6 88*, 6 (6 2015), 1–10. https://link.springer.com/article/10.1140/epjb/e2015-60106-6.

[203] SCHLÖGL, M., AND LEJTOVICZ, K. A Prosopographical Information System (APIS). *ceur-ws.org*. http://ceur-ws.org/Vol-2119/paper9.pdf.

[204] SEMANTIC LAB AT PRATT. Linked Jazz Api. https://linkedjazz.org/api/, Accessed 11 May 2023.

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

[206] SHARMA, D., AND SUROLIA, A. Degree Centrality. *Encyclopedia of Systems Biology* (2013), 558–558. https://link.springer.com/referenceworkentry/10.1007/978-1-4419-9863-7_935.

[207] SPIELMAN, D. Spectral graph theory. *Combinatorial scientific computing 18* (2012), 18.

[208] STANFORD UNIVERSITY. Mapping the Republic of Letters. http://republicofletters.stanford.edu/, Accessed 15 December 2022.

[209] STONE, L. Prosopography. *JSTOR* (1971). https://www.jstor.org/stable/20023990.

[210] SUOMALAINEN KIRJALLISUUDEN SEURA. Finnish Biographies. https://kansallisbiografia.fi/english, Accessed 1 September 2022.

[211] SUOMALAISEN KIRJALLISUUDEN SEURA. Elias Lönnrot Letters. https://www.finlit.fi/en/elias-lonnrot-letters, Accessed 30 September 2019.

[212] SVENSKA LITERATURSÄLLSKAPET I FINLAND. Albert Edelfelts brev. https://edelfelt.fi/, Accessed 7 November 2022.

[213] TAMPER, M., HYVÖNEN, E., AND LESKINEN, P. Visualizing and Analyzing Networks of Named Entities in Biographical Dictionaries for Digital Humanities Research. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICling 2019)* (October 2021), Springer-Verlag. Forth-coming.

[214] TAMPER, M., LEAL, R., SINIKALLIO, L., LESKINEN, P., TUOMINEN, J., AND HYVÖNEN, E. Extracting Knowledge from Parliamentary Debates for Studying Political Culture and Language. In *Proceedings of the 1st International Workshop on Knowledge Graph Generation From Text and the 1st International Workshop on Modular Knowledge co-located with 19th Extended Semantic Conference (ESWC 2022)* (May 2022), S. Tiwari, N. Mihindukulasooriya, F. Osborne, D. Kontokostas, J. D'Souza, and M. Kejriwal, Eds., vol. 3184, CEUR WS, pp. 70–79. International Workshop on Knowledge Graph Generation from Text (TEXT2KG 2022).

[215] TAMPER, M., LESKINEN, P., APAJALAHTI, K., AND HYVÖNEN, E. Using Biographical Texts as Linked Data for Prosopographical Research and Applications. In *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. 7th International Conference, EuroMed 2018, Nicosia, Cyprus* (November 2018), Springer-Verlag.

[216] TAMPER, M., OKSANEN, A., TUOMINEN, J., HIETANEN, A., AND HYVÖNEN, E. Automatic Annotation Service APPI: Named Entity Linking in Legal Domain. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12124 LNCS* (2020), 208–213.

[217] TAMPER, M., SINIKALLIO, L., TUOMINEN, J., AND HYVÖNEN, E. Transforming Linguistically Annotated Finnish Parliamentary Debates Into the Parla-CLARIN Format. In *Digital Humanities in the Nordic and Baltic Countries Seventh Conference (DHNB 2023), Book of Abstracts* (3 2023), S. Gilbert and A. Rockenberger, Eds., University of Oslo Library, Oslo, Norway, p. 118. https://doi.org/10.5281/zenodo.7670464.

[218] TEI CONSORTIUM. Text Encoding Initiative (TEI). https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-correspDesc.html, Accessed 13 January 2023.

[219] TER BRAAKE, S., FOKKENS, A., SLUIJTER, R., DECLERCK, T., AND WANDL-VOGT, E. Biographical Data in a Digital World 2015 : Proceedings of the First Conference on Biographical Data in a Digital World 2015, Amsterdam, The Netherlands, April 9, 2015. *Workshop proceedings 1399* (2015).

[220] THE GENEALOGICAL SOCIETY OF FINLAND. HisKi project - Parish list. https://hiski.genealogia.fi/hiski?en, Accessed 31 March 2022.

[221] THE GETTY RESEARCH INSTITUTE. Getty Union List of Artist Names (Research at the Getty). https://www.getty.edu/research/tools/vocabularies/ulan/, Accessed 1 September 2022.

[222] THE NATIONAL LIBRARY OF FINLAND. About Kanto in English - Toimijakuvailupalvelu - Global Site. https://www.kiwi.fi/display/Toimijakuvailupalvelu/About+Kanto+in+English, Accessed 31 March 2022.

[223] THOMAS, K. Changing Conceptions of National Biography: The Oxford DNB in Historical Perspective. *Changing Conceptions of National Biography: The Oxford DNB in Historical Perspective* (1 2005), 1–56. https://www.cambridge.org/core/books/changing-conceptions-of-national-biography/4AB057D1C9A86E2010088123D52FC232.

[224] THORVALDSEN, G., ANDERSEN, T., AND SOMMERSETH, H. L. Record linkage in the historical population register for Norway. *Population Reconstruction* (1 2015), 155–172. https://doi.org/10.1007/978-3-319-19884-2_8.

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

[226] TUOMINEN, J., HYVÖNEN, E., AND LESKINEN, P. Bio CRM: A data model for representing biographical data for prosopographical research. In *CEUR Workshop Proceedings* (2018), vol. 2119. http://ceur-ws.org/Vol-2119/paper10.pdf.

[227] TUOMINEN, J., KOHO, M., PIKKANEN, I., DROBAC, S., ENQVIST, J., HYVÖNEN, E., LA MELA, M., LESKINEN, P., PALOPOSKI, H.-L., AND RANTALA, H. Constellations of Correspondence: a Linked Data Service and Portal for Studying Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland. In *6th Digital Humanities in Nordic and Baltic Countries Conference, short paper. Accepted for presentation, paper under review* (2022). https://seco.cs.aalto.fi/publications/2022/tuominen-et-al-coco-dhnb-2022.pdf.

[228] TUOMINEN, J., MÄKELÄ, E., HYVÖNEN, E., BOSSE, A., LEWIS, M., AND HOTSON, H. Reassembling the Republic of Letters - A Linked Data Approach. *CEUR Workshop Proceedings 2084* (2018), 76–88. https://helda.helsinki.fi/handle/10138/312719.

[229] TURK, J. jellyfish. https://jamesturk.github.io/jellyfish/, Accessed 10 November 2022.

[230] TURNER, W. Pandasrdf integrates Pandas and RDF. https://github.com/westurner/pandasrdf, Accessed 10 October 2023.

[231] UREÑA-CARRION, J., LESKINEN, P., TUOMINEN, J., VAN DEN HEUVEL, C., HYVÖNEN, E., AND KIVELÄ, M. Communication Now and Then: Analyzing the Republic of Letters as a Communication Network. *Applied Network Science 7* (May 2022). https://doi.org/10.1007/s41109-022-00463-1.

[232] VAN AGGELEN, A., HOLLINK, L., KEMMAN, M., KLEPPE, M., AND BEUNDERS, H. The debates of the European Parliament as linked open data. *Semantic Web, 2017, content.iospress.com* (2016), 1–11. https://content.iospress.com/articles/semantic-web/sw227.

[233] VAN DEN BERG, N., VAN DIJK, I. K., MOURITS, R. J., SLAGBOOM, P. E., JANSSENS, A. A., AND MANDEMAKERS, K. Families in comparison: An individual-level comparison of life-course and family reconstructions between population and vital event registers. *Population Studies 75*, 1 (2021), 91–110.

[234] VAN DEN HEUVEL, C. Mapping Knowledge Exchange in Early Modern Europe: Intellectual and Technological Geographies and Network Representations. *researchgate.net 9*, 1 (3 2015), 95–114. https://www.euppublishing.com/doi/abs/10.3366/ijhac.2015.0140?journalCode=ijhac.

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

[236] VAN MIERT, D. What was the Republic of Letters? A brief introduction to a long history (1417–2008). *Groniek 204/205* (11 2016), 269–287.

[237] VENTURINI, T., OUESTWARE, AND CNRS CIS. Retina. https://ouestware.gitlab.io/retina/1.0.0-beta.1, Accessed 14 November 2023.

[238] VERBOVEN, K., CARLIER, M., AND DUMOLYN, J. A Short Manual to the Art of Prosopography. In *Prosopography Approaches and Applications. A Handbook*, K. S. B. Keats-Rohan, Ed. Unit for Prosopographical Research (Linacre College), 2007, pp. 35–70. http://dx.doi.org/1854/8212.

[239] VESPIGNANI, A. Twenty years of network science. *Nature 558*, 7711 (6 2018), 528–529.

[240] VIRTANEN, A., KANERVA, J., ILO, R., LUOMA, J., LUOTOLAHTI, J., SALAKOSKI, T., GINTER, F., AND PYYSALO, S. Multilingual is not enough: BERT for Finnish. https://arxiv.org/abs/1912.07076v1.

[241] VRANDEČIĆ, D., AND KRÖTZSCH, M. Wikidata: A Free Collaborative Knowledgebase. *scholar.archive.org 57*, 10 (9 2014), 78–85. https://dl.acm.org/doi/pdf/10.1145/2629489.

[242] WAGNER, R. A., AND FISCHER, M. J. The String-to-String Correction Problem. *Journal of the ACM (JACM) 21*, 1 (1 1974), 168–173.

[243] 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* (2018). https://doi.org/10.22148/16.028.

[244] WARREN, C. N., OTIS, J., WANG, L., FINEGOLD, M., AND SHALIZI, C. Six degrees of Francis Bacon: A statistical method for reconstructing large historical social networks. *hcommons.org* (2016). https://hcommons.org/deposits/item/mla:989/.

[245] WASSERMAN, S., AND FAUST, K. Social network analysis: Methods and applications. *Cambridge university press* (1994). https://toc.library.ethz.ch/objects/pdf_ead50/3/E28_1502716_TB-I_002336476.pdf.

[246] WATTS, D. J., AND STROGATZ, S. H. Collective dynamics of 'small-world' networks. *nature.com*. https://www.nature.com/articles/30918., Accessed 14 January 2022.

[247] WICKHAM, H., ÇETINKAYA-RUNDEL, M., AND GROLEMUND, G. *R for data science*. O'Reilly Media, Inc., 2023.

[248] WIKIMEDIA FOUNDATION. Wikidata. https://www.wikidata.org/wiki/Wikidata:Main_Page, Accessed 1 September 2023.

[249] WIKIMEDIA FOUNDATION. Wikidata: SPARQL query service/queries/examples. https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples/en, Accessed 1 September 2023.

[250] WILSON, P. Collaborative knowledge building: Ethnographic insights from Geni.com. *ieeexplore.ieee.org* (2011), 999–1007. https://ieeexplore.ieee.org/abstract/document/5967202/.

[251] WINDHAGER, F., FEDERICO, P., SCHREDER, G., GLINKA, K., DORK, M., MIKSCH, S., AND MAYR, E. Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. *IEEE Transactions on Visualization and Computer Graphics 25*, 6 (6 2019), 2311–2330.

[252] WINKLER, W. E. Overview of record linkage and current research directions. *Bureau of the Census 25*, 4 (2006), 603–623. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.79.1519.

[253] WORLD WIDE WEB CONSORTIUM. Semantic Web - W3C. https://www.w3.org/standards/semanticweb/, Accessed 5 September 2022.

[254] WYNAR, B. S., TAYLOR, A. G., AND OSBORN, J. Introduction to cataloging and classification. *Libraries unlimited Littleton, CO 8*, 2 (1980). https://www.academia.edu/download/3445870/202-20060914.pdf.

[255] YAN, E., AND DING, Y. Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology 63*, 7 (7 2012), 1313–1326. https://onlinelibrary.wiley.com/doi/full/10.1002/asi.22680.

Publication I

Petri Leskinen, Mikko Koho, Erkki Heino, Minna Tamper, Esko Ikkala, Jouni Tuominen, Eetu Mäkelä, and Eero Hyvönen. Modeling and Using an Actor Ontology of Second World War Military Units and Personnel. In *The Semantic Web – ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part II*, Claudia d'Amato, Miriam Fernandez, Valentina Tamma, Freddy Lecue, Philippe Cudré-Mauroux, Juan Sequeda, Christoph Lange, and Jeff Heflin (editors), Information Systems and Applications, incl. Internet/Web, and HCI, volume 10588, pages 280–296, ISBN 9783319682037, Springer, Cham, October 2017, online https://doi.org/10.1007/978-3-319-68204-4_27.

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

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

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

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

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

1 Introduction

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

[3] http://www.getty.edu/research/tools/vocabularies/ulan/about.html
[4] http://xmlns.com/foaf/spec/
[5] http://vocab.org/relationship/
[6] http://vocab.org/bio/

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

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

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

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

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

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

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

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

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

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

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

2 Use Case and Datasets

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

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

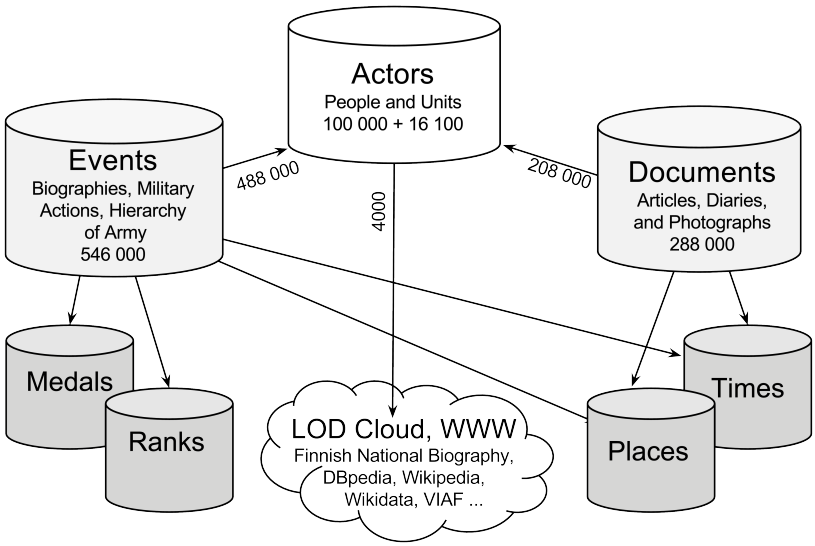


Fig. 1. Linkage in the actor-event based dataset

3 Actor Ontology Model

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

[11] http://sotasampo.fi/en/
[12] http://cidoc-crm.org/

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

Table 1. Namespaces and prefixes used in actor ontologies

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

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

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

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

[13] `http://www.cidoc-crm.org/cidoc-entities/e39-actor`
[14] `http://www.cidoc-crm.org/cidoc-entities/e5-event`
[15] `http://ldf.fi/warsa/actors/person_294.ttl`

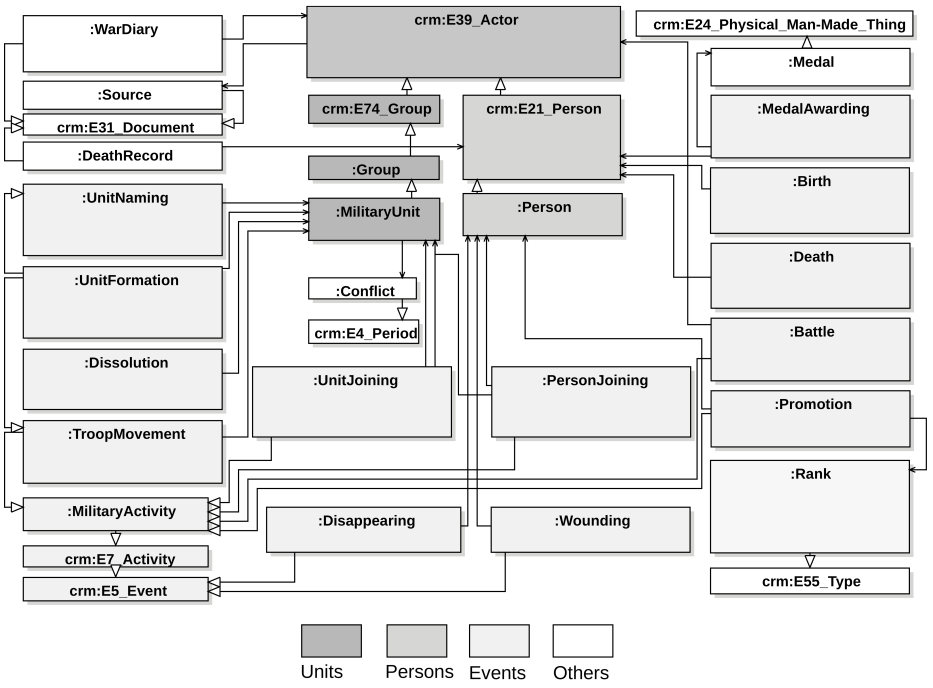


Fig. 2. Ontology schema of actors and events

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

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

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

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

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

Table 2. Properties of a resource describing pilot Jorma Karhunen

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

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

4 Warsampo Actor Data

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

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

[17] kronos.narc.fi/menehtyneet/
[18] http://www.kansallisbiografia.fi/english/
[19] http://sa-kuva.fi/neo?tem=webneoeng
[20] https://www.wikidata.org/

Table 3. Examples of events describing pilot Jorma Karhunen

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

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

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

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

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

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

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

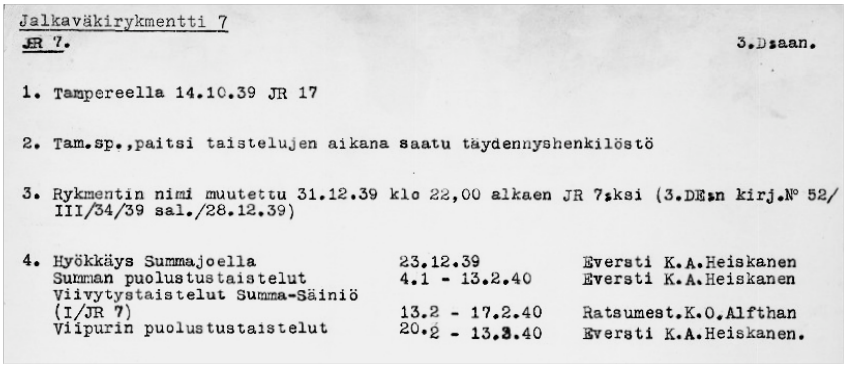


Fig. 3. Information on an Organization Card

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

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

Table 5. Examples of events describing 24th Fighter Squadron

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

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

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

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

[22] `http://wiki.dbpedia.org/`
[23] `https://viaf.org/`
[24] `https://en.wikipedia.org/w/api.php`
[25] `https://pypi.python.org/pypi/wikipedia`
[26] See, e.g., `http://ldf.fi/warsa/conflicts/LaplandWar`
[27] See, e.g., `http://digi.narc.fi/digi/hae_ay.ka?sartun=319.SARK`

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

5 The WarSampo portal

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

5.1 The Person Perspective

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

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

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

[28] See, e.g., `http://ldf.fi/warsa/medals/medal_83`
[29] See, e.g., `http://ldf.fi/warsa/actors/ranks/Majuri`
[30] `http://www.sotasampo.fi/en/`
[31] `http://www.sotasampo.fi/en/persons`
[32] `http://yasgui.org/short/B1w2O71gb`
[33] `http://sotasampo.fi/en/units`

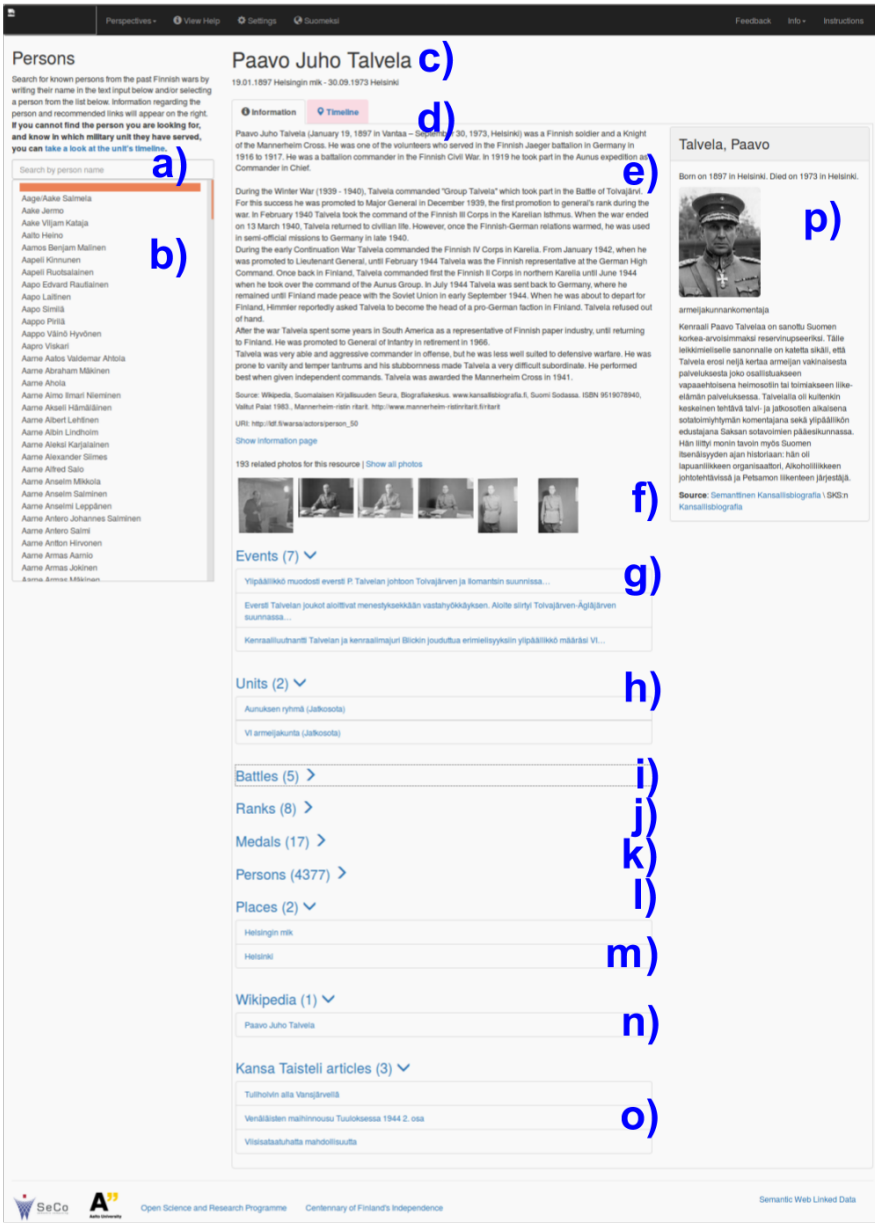


Fig. 4. Information on Person Perspective

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

5.2 The Military Unit Perspective

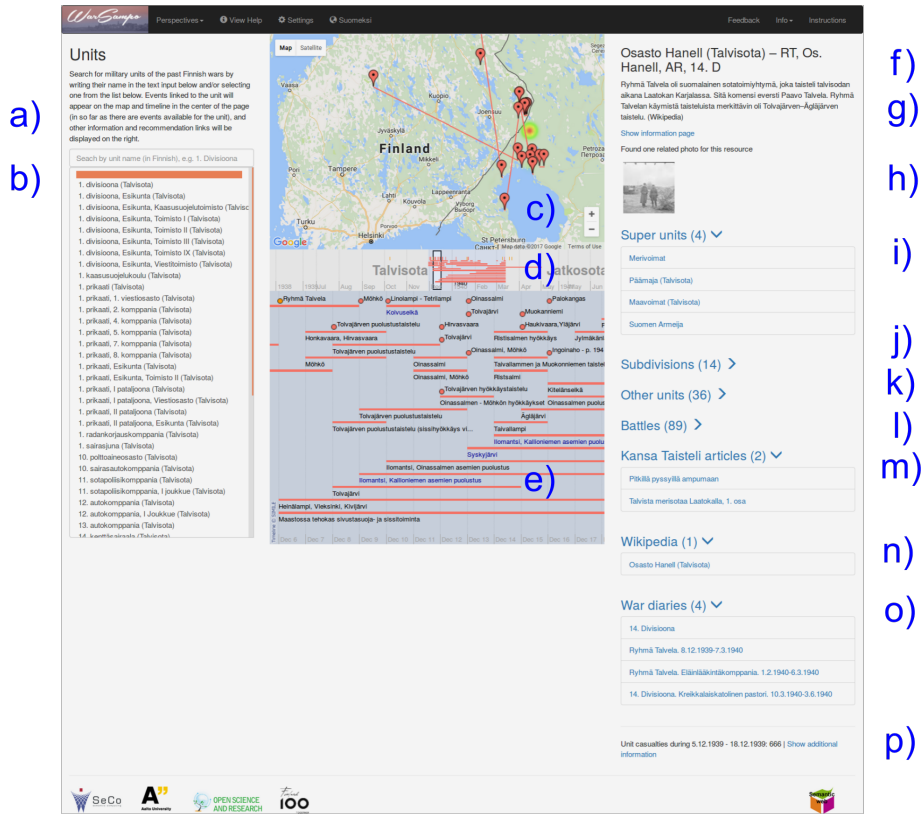


Fig. 5. Information on Unit Perspective

5.3 The Kansa Taisteli Magazine Perspective

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

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

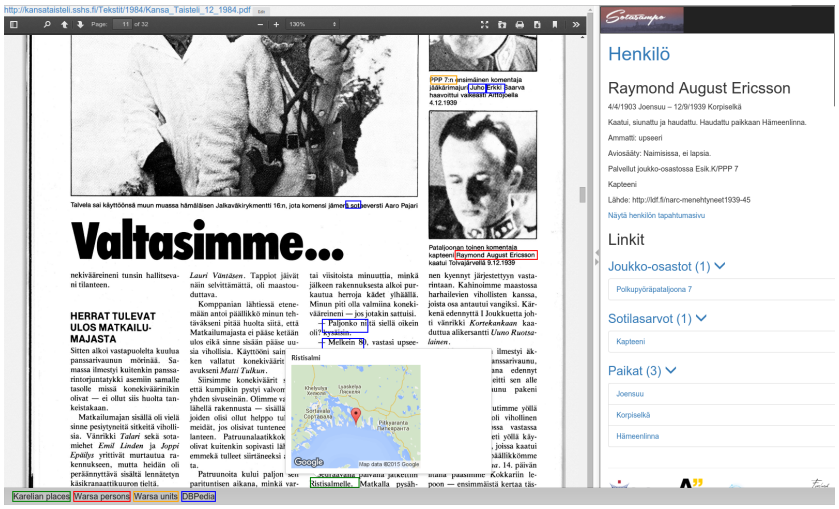


Fig. 6. The Contextual Reader interface targeting the Kansa Taisteli magazine articles

The purpose of the perspective is two-fold: 1) to help a user find articles of interest using faceted semantic search and, 2) to provide context to the found articles by extracting links to related WarSampo data from the texts. The start page of the magazine article perspective is a faceted search browser. Here, the facets allow the user to find articles by filtering them based on author, issue, year, related place, army unit, or keyword. Some of the underlying properties, such as the year and issue number, are hierarchical and represented using SKOS. The hierarchy is visualized in the appropriate facet, and can be used for query expansion: by selecting an upper category in the facet hierarchy one can perform a search using all subcategories.

After the user has found an article of interest, she can click on it, and the article appears on the screen in the CORE Contextual Reader interface [14]. Depicted in Fig. 6, CORE is able to automatically and in real time annotate PDF and HTML documents with recognized keywords and named entities, such as army units, places, and person names. These are then encircled with colored boxes indicating the linked data source. By hovering the mouse over a box, data is shown to the user, providing contextual information for an enhanced reading experience. In Fig. 6, for example, detailed data are shown about *Raymond August Ericsson*, one of the battalion commanders discussed in the article.

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

based on the war into the corresponding units. A detailed description and evaluation of the process is available at [17].

5.4 Photographs

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

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

6 Related work, and Discussion

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

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

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

[35] http://www.sotasampo.fi/en/photographs
[36] http://www.europeana-collections-1914-1918.eu
[37] http://www.1914-1918-online.net
[38] http://ww1.discovery.ac.uk
[39] http://www.canadiana.ca/en/pcdhn-lod/
[40] http://thesaurus.historicengland.org.uk/thesaurus.asp?thes_no=365&thes_name=Defence%20of%20Britain%20Thesaurus
[41] http://www.bygle.net/wp-content/uploads/2015/04/Open-Memory-Project_3-1.pdf
[42] http://seco.cs.aalto.fi/applications/kulttuurisampo/
[43] http://seco.cs.aalto.fi/applications/kirjasampo/
[44] http://seco.cs.aalto.fi/projects/biographies/

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

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

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

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

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

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

References

1. de Boer, V., van Doornik, J., Buitinck, L., Marx, M., Veken, T.: Linking the kingdom: enriched access to a historiographical text. In: Proceedings of the 7th International Conference on Knowledge Capture (KCAP 2013). pp. 17–24. ACM (June 2013)
2. Bunescu, R.C., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: EACL. vol. 6, pp. 9–16 (2006)
3. Collins, T., Mulholland, P., Zdrahal, Z.: Semantic browsing of digital collections. In: Proceedings of the 4th International Semantic Web Conference (ISWC 2005). pp. 127–141. Springer–Verlag (November 2005)
4. Doerr, M.: The CIDOC CRM – an ontological approach to semantic interoperability of metadata. AI Magazine 24(3), 75–92 (2003)

[45] http://openscience.fi/

5. Guha, R.V., Brickley, D., Macbeth, S.: Schema. org: Evolution of structured data on the web. Communications of the ACM 59(2), 44–51 (2016)
6. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R.: Evaluating entity linking with Wikipedia. Artificial Intelligence 194, 130–150 (Jan 2013), http://dx.doi.org/10.1016/j.artint.2012.04.005
7. Heino, E., Tamper, M., Mäkelä, E., Leskinen, P., Ikkala, E., Tuominen, J., Koho, M., Hyvönen, E.: Named Entity Linking in a Complex Domain: Case Second World War History. In: Language, Technology and Knowledge 2017. June 19-20, Galway, Ireland. Springer-Verlag (2017), in press
8. Hyvönen, E., Ikkala, E., Tuominen, J.: Linked data brokering service for historical places and maps. In: Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe). pp. 39–52. CEUR Workshop Proceedings (May 2016), http://ceur-ws.org/Vol-1608/#paper-06, vol 1608
9. Hyvönen, E.: Cultural heritage linked data on the semantic web: Three case studies using the sampo model (2017), http://seco.cs.aalto.fi/publications/submitted/hyvonen-vitoria-2017.pdf, invited talk, Proceedings of the VIII Encounter of Documentation Centres of Contemporary Art: Open Linked Data and Integral Management of Information in Cultural Centres Artium, Vitoria-Gasteiz, Spain, 2016. Forthcoming
10. Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J., Mäkelä, E.: WarSampo Data Service and Semantic Portal for Publishing Linked Open Data about the Second World War History. In: The Semantic Web – Latest Advances and New Domains (ESWC 2016). pp. 758–773. Springer-Verlag (2016)
11. Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., Sirola, L.: Reassembling and enriching the life stories in printed biographical registers: Norssi high school alumni on the semantic web. In: Proceedings, Language, Technology and Knowledge 2017. June 19-20, Galway, Ireland. Springer-Verlag (February 2017), http://ldk2017.org/, accepted
12. Koho, M., Hyvönen, E., Heino, E., Tuominen, J., Leskinen, P., Mäkelä, E.: Linked death - representing, publishing, and using second world war death records as linked open data. In: Sack, H., Rizzo, G., Steinmetz, N., Mladenić, D., Auer, S., Lange, C. (eds.) The Semantic Web: ESWC 2016 Satellite Events. Springer-Verlag (June 2016)
13. Mäkelä, E.: Combining a rest lexical analysis web service with sparql for mashup semantic annotation from text. In: European Semantic Web Conference. pp. 424–428. Springer (2014)
14. Mäkelä, E., Lindquist, T., Hyvönen, E.: CORE - a contextual reader based on linked data. In: Proceedings of Digital Humanities 2016, long papers (July 2016)
15. Mäkelä, E., Törnroos, J., Lindquist, T., Hyvönen, E.: WW1LOD - An application of CIDOC-CRM to World War 1 Linked Data. International Journal on Digital Libraries (2016), in press.
16. Szekely, P., Knoblock, C.A., Yang, F., Zhu, X., Fink, E.E., Allen, R., Goodlander, G.: Connecting the Smithsonian American Art Museum to the Linked Data Cloud. In: Extended Semantic Web Conference. pp. 593–607. Springer (2013)
17. Tamper, M., Leskinen, P., Ikkala, E., Oksanen, A., Mäkelä, E., Heino, E., Tuominen, J., Koho, M., Hyvönen, E.: AATOS – a Configurable Tool for Automatic Annotation. In: Proceedings, Language, Technology and Knowledge 2017. June 19-20, Galway, Ireland. Springer-Verlag (February 2017), accepted
18. Taylor, A.: Introduction to cataloging and classification. Library and Information Science Text Series, Libraries Unlimited (2006)
19. Warren, R.: Creating specialized ontologies using Wikipedia: The Muninn experience. Berlin, DE: Proceedings of Wikipedia Academy: Research and Free Knowledge (WPAC2012). URL: http://hangingtogether. org (2012)

Publication II

Petri Leskinen, Jouni Tuominen, Erkki Heino, and Eero Hyvönen. An Ontology and Data Infrastructure for Publishing and Using Biographical Linked Data. In *Proceedings of the Workshop on Humanities in the Semantic Web (WHiSe II)*, Alessandro Adamou, Enrico Daga, Leif Isaksen (editors), CEUR Workshop Proceedings, pages 15-26, Vienna, Austria, October, 2017, online https://ceur-ws.org/Vol-2014/paper-02.pdf.

An Ontology and Data Infrastructure for Publishing and Using Biographical Linked Data

Petri Leskinen[1], Jouni Tuominen[1,2], Erkki Heino[1,2], and Eero Hyvönen[1,2]

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

Abstract. This paper describes the ontology model and published datasets of a digitized biographical person register. The applied ontology model is designed to represent people via their enduring roles and perduring lifetime events. The model is designed to support 1) prosopographical Digital Humanities research, 2) linking to resources in semantic Cultural Heritage portals, and 3) semantic data validation and enrichment by using SPARQL queries. The linked data approach enables to enrich a person's biography by interlinking it with space and time related biographical events, persons relating by social contacts or family relations, historical events, and personal achievements.

Keywords: Semantic Web, Linked Open Data, Actor Ontology, Digital Humanities, Cultural Heritage, Prosopography, Biographical Representation

1 Introduction

This resource description paper presents the LOD infrastructure, data model, and datasets used in the Norssit alumni register of short biographies. The data model is designed to support prosopographical research, data aggregation and linking in semantic portals, and easy SPARQL querying. The datasets consist of 10 137 person resources, enriched with graphs of relating career events and family relations, and vocabularies of titles, schools, companies, medals, and hobbies.

The data has been used in creating the Vanhat Norssit Portal[3] allowing the user to search and browse the data about individual persons as well as analyze and visualize data about groups of people in proposographical research [1,11]. The user can filter the results by making selections on the facets on the left side of the page. For visualizing the data the portal has two views that use Google Chart[4] diagrams. On the first one[5], the pie charts show the popularity of most common educations, universities and colleges, professions, and employers after the graduation of the alumni. On the second one[6], years of enrollment and matriculation are shown using histograms, and below these,

[3] http://www.norssit.fi/semweb
[4] https://developers.google.com/chart/
[5] http://www.norssit.fi/semweb/#!/visualisointi
[6] http://www.norssit.fi/semweb/#!/visualisointi2

multi-column charts showing the most popular universities and colleges, employers, and occupations by decade. The digitization, lodification, and the Vanhat Norssi Portal is described in more details in [8].

This paper is structured as follows: First, an ontology model for representing people with their life time events and relation roles is introduced. Secondly, the data sets of the use case Norssit alumni with information extraction from textual data is discussed. Then the results of entity linking are evaluated. Finally, the related work, lessons learned, and future work are discussed.

2 Person Ontology Model

The ontology model representing people and their biographical information in the use case Norssit alumni is based on the Bio CRM model[7], which has been developed to facilitate and harmonize the representation of biographies and cultural heritage data on the Semantic Web. Bio CRM is a domain specific extension of CIDOC CRM[8] [3]. It includes structures for basic data of people, personal relations, professions, and events with participants in different roles. Bio CRM makes a distinction between enduring unary roles of actors, their enduring binary relationships, and perduring events, where the participants can take different roles modeled as a role concept hierarchy.[9] Bio CRM provides the general data model for biographical datasets, and the individual datasets concerning different cultures, time periods, or collected by different researchers may introduce extensions for defining additional event and role types.

Namespace	**Prefix**
http://ldf.fi/norssit/	:
http://ldf.fi/schema/bioc/	bioc:
http://purl.org/dc/terms/	dct:
http://xmlns.com/foaf/0.1/	foaf:
http://schema.org/	schema:
http://www.w3.org/2004/02/skos/core#	skos:
http://www.w3.org/1999/02/22-rdf-syntax-ns#	rdf:
http://www.w3.org/2000/01/rdf-schema#	rdfs:

Table 1. The namespaces and prefixes used in the ontology.

The main classes of the person ontology are shown in Figure 1. A principle is that a **foaf:Person** instance only contains the properties that are considered constant, in our case family and first names, places, and dates of birth and death, etc. The person resource is enriched by attaching family relations, achievements, and titles.

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

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

[9] http://ldf.fi/schema/bioc/

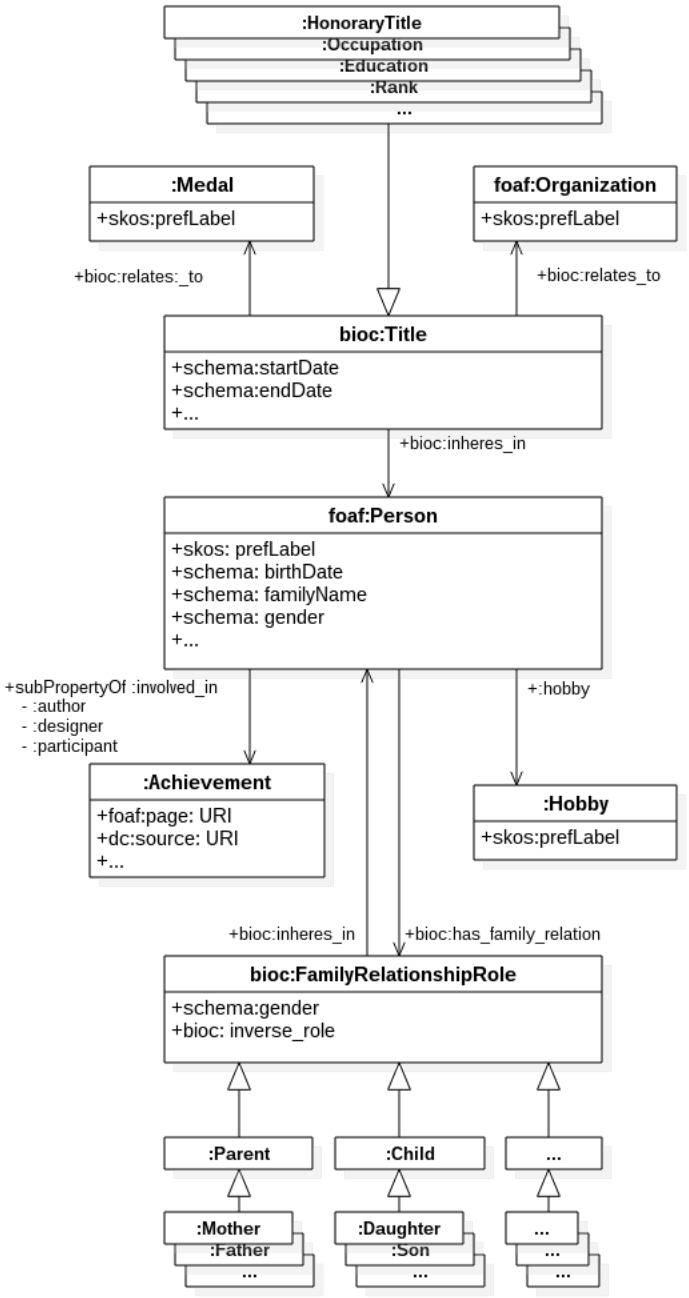


Fig. 1. The ontology schema.

2.1 Modeling Family Relations

Family relation is an example of a binary, often even N-nary, relationship connecting two or more people. Social relationships can be modeled in a similar manner. Each family relation is a subclass of **bioc:FamilyRelationshipRole**. The domain specific ontology of family relations can build a hierarchy (e.g. **:Mother** and **:Father** are subclasses of **:Parent**). The RDF example below shows a definition of a class and declaration of a relationship between two people. Notice that the property **bioc:inverse_role** has two values, depending on the gender of the relative. Genders and inverse relations can be used for data evaluation and reasoning: this SPARQL query[10] fills in the missing inverse relationships in the dataset. A family relation is attached to a person using the property **bioc:has_family_relation**, which can have a blank node or a resource as a value.

```
:Mother  a                        rdfs:Class ;
        rdfs:subClassOf           :Parent ;
        bioc:inverse_role         :Son , :Daughter ;
        schema:gender             schema:Female ;
        skos:prefLabel            "Mother"@en , "äiti"@fi .

:person_1  a                      foaf:Person ;
        schema:gender             schema:Male ;
        bioc:has_family_relation    [ a :Mother ;
                                      bioc:inheres_in :mother_1 ] .
```

2.2 Modeling Personal Achievements

A personal achievement refers to any notable activity (e.g. producing a work of art, a design project, receiving a political achievement, or participating in athletes games). Achievements are modeled so that a subproperty of **:involved_in** connects the person to a instance of class **:Achievement**, and furthermore indicates what is the role of the person in a particular achievement. So a single achievement can refer to multiple people, and indicate the roles they participated with, e.g. an actor, author, or director of a movie. Instances of **:Achievement** can contain a description, information of time, place, and provenance, and link to a corresponding LOD resource or to a web page.

```
:norssi_2230  :author  :achievement_33 .

:author a                   rdf:Property ;
        skos:prefLabel      "teoksia"@fi , "Novels"@en ;
        rdfs:subPropertyOf  :involved_in .

:achievement_33  a          :Achievement ;
        skos:prefLabel      "Sinuhe egyptiläinen."@fi ,
                            "Sinuhe the Egyptian."@en ;
        dct:source          <https://fi.wikipedia.org/wiki?curid=820> ;
        :wikipedia
        <https://en.wikipedia.org/wiki/The_Egyptian> .
```

[10] Example of a SPARQL query: `http://yasgui.org/short/BkRKKXyIZ`

2.3 Modeling Career Roles

According to the Bio CRM principle, the occupation or profession of a person is considered a role. Another unary role modeled in a similar manner would be person's nationality. The resource is a subclass of **bioc:Title**. The person involved is attached with a subproperty of **bioc:inheres_in**, additional information like the company or medal of honour by subproperties of **bioc:relates_to**.

```
:tekniikan%20tohtori
        a                   rdfs:Class ;
        rdfs:subClassOf     :Education ;
        skos:altLabel       "TkT"@fi, "D.Sc. (Tech.)"@en ;
        skos:prefLabel      "tekniikan tohtori"@fi ,
                            "Doctor of Science (Technology)"@en .

:event_21721  a             :tekniikan%20tohtori ;
        bioc:inheres_in     :norssi_7686 ;
        schema:startDate    "1991"^^xsd:gYear ;
        skos:prefLabel      "TkT 91"@fi .
```

3 Norssit Dataset

As a concrete case study, a register 1867–1992 of over 10 000 alumni of the prominent Finnish high school "Norssi" was scanned, OCR'd, and transformed into RDF, then enriched by data linking, published as a linked data service, and finally provided to end users via a faceted search engine and browser for studying lives of historical persons and for prosopographical research. [8]

3.1 Information Extraction

The most important data source was the textual description of register entries. In Figure 2 a register entry is depicted, and some of the data fields are picked as examples. Description texts are well-formated, and always start with person's name (a) and his place and date of birth (b, Jyväskylä, 19th Aug., 1868). Description includes names of person's parents, and his years of enrollment and matriculation (c, *yo 88*). His later university degrees with graduation years are mentioned (d, *FT = Ph.D.*). The career is described as a list of entries in format *Company role years* (e, *toimJ = CEO*). Possible medals of honour are mentioned (f, *VirVR 1 mk = Estonian Cross of Liberty, 1st Class*) as well as military ranks with promotion years (g, *Evl = Colonel Lieutenant*). The description ends with a possible date of death (h, *11th Dec., 1939*) and list of family relations (i, *Veli = Brother)*. The data fields were extracted using regular expressions.

3.2 Datasets

Currently the Norssit dataset contains ca 892 000 triples defining 131 000 resources. The main graphs are discussed in detail below, with the graph names, amounts of instances and triples, main classes, and properties.

Fig. 2. A biographical entry in the register book with examples of extracted data fields.

People

Graph:	`http://ldf.fi/norssit/people`
Contains:	17 791 instances, 183 972 triples
Core classes:	**foaf:Person**
Properties:	**schema:familyName**, **schema:givenName**, **schema:gender**, **skos:prefLabel**, **schema:birthDate**, **schema:birthPlace**

This graph includes data of 10 137 people. A person resource contains biographical data extracted from register descriptions, e.g. given and family names, gender, places and dates of birth and death, years of enrollment, matriculation, or resignation, and provenance data. This graph also includes profile image URIs, family relationship instances, and links to external LOD clouds.

Career events

Graph:	`http://ldf.fi/norssit/events`
Contains:	34 000 instances, 247 237 triples
Core classes:	subclasses of **bioc:Title**
Properties:	**bioc:inheres_in**, **bioc:relates_to**, **schema:startDate**, **schema:endDate**

The event graph contains 34 000 career events extracted from register descriptions (e.g. see e in Fig. 2). Each resource contains links to an occupational or educational title, person involved, start and end years, a description, and a possible organization or medal. Altogether 5882 people are enriched with a title.

Achievements

Graph:	`http://ldf.fi/norssit/achievements`
Contains:	3000 instances, 15 578 triples
Core class:	**:Achievement**
Properties:	**:involved_in**, **skos:prefLabel**, **dct:source**, **:wikipedia**

The achievement graph contains 3000 personal achievements extracted from Wikipedia pages, or BookSampo[11] Linked Data. In the case of Wikipedia, the information was extracted from HTML code under specified subtitles. Each resource provides a description, specified category, links to a Wikipedia page of the achievement, and link to person's Wikipedia page served as an source of information.

Organizations

Graph:	`http://ldf.fi/norssit/organizations`
Contains:	2300 instances, 5390 triples
Core classes:	**foaf:Organization**, **schema:EducationalOrganization**
Properties:	**skos:prefLabel**, **skos:altLabel**, **dct:source**

The organization graph contains the labels and abbreviations of 2401 organizations, e.g. government agencies, companies, colleges, or universities. The labels were collected from text descriptions. An organization is attached to an event using the **bioc:relates_to** property. Altogether 4805 people are linked to an organization.

Hobbies

Graph:	`http://ldf.fi/norssit/hobbies`
Contains:	1760 instances, 3520 triples
Core class:	**:Hobby**
Property:	**skos:prefLabel**

The vocabulary of hobbies contains labels of 1760 different hobbies (e.g. Music, Sports, or Arts), mentioned in register descriptions. A hobby is attached using the **:hobby** property of Person resource. Altogether 7845 people are related with a hobby.

Titles

Graph:	`http://ldf.fi/norssit/titles`
Contains:	350 instances, 1526 triples
Core classes:	subclasses of **bioc:Title**
Properties:	**skos:prefLabel**, **skos:altLabel**

The titles graph contains classes of 350 educational or occupational titles and military ranks. These are the classes of instances in Career events graph. Altogether 5882 people have a specified title.

Medals

Graph:	`http://ldf.fi/norssit/medals`
Contains:	301 instances, 1254 triples
Core class:	**:Medal**
Property:	**skos:prefLabel**

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

The medals vocabulary contains 301 different types of honorary medals, the data is extracted from register descriptions (e.g. see f in Fig. 2). A medal is attached to an instance in Career events graph using the **:relates_to_medal** property. Altogether 1844 people are mentioned with a medal.

Bio CRM schema

Graph:	`http://ldf.fi/schema/bioc`
Contains:	451 triples
Core classes:	subclasses of **bioc:FamilyRelationshipRole**
Properties:	**bioc:inheres_in**, **:inverse_role**, **schema:gender**

This graph contains the definitions of the core classes and properties of the Bio CRM schema. It also includes the domain specific definitions of 45 subclasses of **bioc:FamilyRelationshipRole** (e.g. family members **:Child**, **:Mother**, or **:Father**) and 24 subproperties of **:involved_in** (e.g. publication, design project, or nomination).

3.3 Linkage to LOD cloud

The Norssit data is linked to external data clouds shown in Table 2. The linkage was done with string comparison using person's full name with known dates of birth and death. Links were created to Wikipedia, Wikidata, National Biography of Finland[12] and its Swedish complement BLF[13], BookSampo Linked Data, CultureSampo[14] portal, WarSampo[15] portal, ULAN[16] authority register by The J. Paul Getty Trust, VIAF[17], and the genealogical data service Geni[18]. For entity linking to databases offering a SPARQL endpoint, the tool SPARQL ARPA[19] was used. In cases where the database provides a REST API, like Wikipedia or Geni.com, a special Python script was used. A Python script was used also in the case of BLF, where the data was only available as a CSV formatted table. [8]

4 Evaluation

The results of information extraction, and external data linking are evaluated in this chapter. Evaluation was done by first choosing an random sample at size of $N = 50$ or $N = 100$ people, and then manually checking if the data extracted or linkage accomplished was correct.

The results in Table 3 are all literal values of such properties that each person should have. Results indicate whether the information was interpreted correctly or not. For the

[12] `http://www.kansallisbiografia.fi/english`
[13] `http://www.sls.fi/sv/projekt/blf-biografiskt-lexikon-finland`
[14] `http://www.kulttuurisampo.fi`
[15] `http://sotasampo.fi/en/`
[16] `http://www.getty.edu/research/tools/vocabularies/ulan/`
[17] `http://www.viaf.org`
[18] `http://www.geni.com`
[19] `http://seco.cs.aalto.fi/projects/dcert/`

Data Source	Links	Description
Geni	894	Family research and family tree data
Wikipedia	609	`http://fi.wikipedia.org`
Wikidata	609	`http://www.wikidata.org`
CultureSampo	453	LOD from museums, archives, libraries, and media
WarSampo	352	Second World War LOD service and portal
National Biography	136	National Biography of Finland
VIAF	135	Virtual International Authority Files
BookSampo	90	Finnish fiction literature on the Semantic Web service
BLF	44	Biografiskt Lexikon för Finland
ULAN	16	Union List of Artist Names Online

Table 2. The data sources linked to the Norssit register.

property of name, the single false result was caused by an error in the OCR process, which caused an erroneous family name in the dataset. In the register book, the dates of birth and death were written in format *dd MM yyyy* with the month in Roman numerals (see b and h in Fig. 2). The two false results were caused by a typical OCR problem of mixing up characters *l*, *1*, and *I*. The year of enrollment was annotated in two digit decade and year format (c in Fig. 2, e.g. *83* for *1883* or *1983*), and the century was automatically reasoned based on the person's birth year.

The Norssi high school has had female pupils only after the year 1972; approximately 11 per cent of people in the data set are female. For pupils enrolled 1972 or after, the gender was generated in three steps. First, depending on the known family relations, some people were marked male or female. Next, given names of people remaining without a specified gender were compared to the given names of people with known gender, and people with matching names inherited the corresponding gender. Finally, a list with less than 100 otherwise unidentified rare names was filled manually.

Description	Property	Correct	False	Precision
Name	skos:prefLabel	49	1	0.98
Date of birth	schema:birthDate	48	2	0.96
Year of enrollment	:enrollmentYear	50	0	1.00
Gender	schema:gender	50	0	1.00

Table 3. Examples of the precision of the text retrieval.

Unlike in the previous examples, the properties evaluated in Table 4 were not obligatory. So, like for instance for the date of death there are 10 true positive (TP) matches where the information was interpreted correctly, 38 true negative (TN) cases where

the data had no such a date, one false positive (FP) case caused by noise in the OCR-process, and one false negative (FN) where the information was not retrieved. The year of matriculation was an integer extracted from the text just like the year of enrollment. Profile images for each corresponding person (see e.g. Fig. 2) were located from the OCR'd layout in the XML-format.

Some of the properties were very sparse in data, so the sample size was increased to $N = 100$ for the evaluation of hobbies, family relations, careers, and links to LOD cloud. The false results in the family relations were caused by misinterpreting certain words of the Finnish language (e.g. the word *Eno* (Uncle) is also a name of a village). The data of the family relationships was further evaluated with SPARQL queries checking some basic rules, e.g. a parent must be older than the child[20].

The algorithm for linking to external databases (Wikipedia, Semantic National Biography of Finland [6], WarSampo [5], and Geni.com) was designed to prefer precision on the expense of a lower recall; e.g. to link entities only in cases when assured that they represent the same person. This linkage could not be done based on person's name solely, and required extra information like places and times of birth and death.

Property	TP	FP	TN	FN	Precision	Recall	F_1 score
Day of death	10	1	38	1	0.91	0.91	0.91
Year of matriculation	15	2	29	4	0.88	0.79	0.83
Profile Image	34	0	63	3	1.00	0.92	0.96
Hobbies	98	0	0	2	1.00	0.98	0.99
Family relations	98	2	0	0	0.98	1.00	0.99
Careers	150	0	0	19	1.00	0.89	0.94
National Biography of Finland	4	0	96	0	1.00	1.00	1.00
WarSampo	3	1	94	2	0.75	0.60	0.67
Wikipedia	5	0	91	4	1.00	0.56	0.71
Geni.com	12	0	81	7	1.00	0.63	0.77

Table 4. Examples of the precision and recall of the dataset linking.

5 Discussion

5.1 Related Work

Our research group, Semantic Computing Research Group (SeCo) has produced several projects with actor ontologies: The National Biography of Finland, CultureSampo[21],

[20] Example of a SPARQL query: `http://yasgui.org/short/rJI8CXyUb`

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

BookSampo[22], and WarSampo [7] datasets, which are all highly interlinked; and linked to the Norssit project as well. The source material in our project was in a clearly structured format, while *Van de Camp* [2] deals with information extraction from unstructured text. *Szekely et al.* [9] describe linking datasets of Smithsonian American Art Museum with DBpedia and the Getty Vocabularies.

CIDOC CRM includes a mechanism for representing the role of an active participant in an event, modeling it as a property of the property that states the participant (see CIDOC's **P14.1 in the role of**). There is a proposal for encoding CIDOC's properties of properties as RDF[23], introducing new class for the property and auxiliary properties, which adds complexity to the data model. Simple Event Model (SEM) [4] is a general model for expressing events, with support for three alternative representations for roles, based on using a) **rdf:value**, b) reification, or c) named graphs. Standards for Networking Ancient Prosopographies (SNAP) project[24] has developed an ontology for representing personal relationships.

Bio CRM's approach aims for simplicity and compatibility with CIDOC CRM. The model supports expressing enduring unary roles and binary relationships without the need to model them in the context of an event.

5.2 Lessons Learned

In our dataset, some practices were simplified, like modeling of a person's birth with literal values of properties **:birth_place** and **:birth_time** instead of using the CIDOC CRM event **crm:E69_Birth**. These simplifications worked well in this case, and reduced the complexity of the data. Another similar case was modeling people's names as literal values instead of using a resource of the type **crm:E41_Appellation**.

5.3 Future Work

We will continue our work on modeling and publishing biographies with data publications dealing with Finnish Biographies[25], and U.S. Congress Legislators[26].

HISCO[27] [10] is a vocabulary of historical occupations and professions, which has a hierarchical structure. We are extending the HISCO vocabulary with Finnish labels, mostly extracted from Wikidata and WordNet[28], some manually translated.

Acknowledgements Our work is part of the project Semantic Web Publications – Texts as Data Services (Severi)[29], funded mainly by Tekes. The development of Bio CRM

[22] http://seco.cs.aalto.fi/applications/kirjasampo/
[23] http://www.cidoc-crm.org/roles-in-the-cidoc%E2%80%90crm-modelling-properties-of-properties
[24] https://snapdrgn.net
[25] http://www.kansallisbiografia.fi/english
[26] https://github.com/unitedstates/congress-legislators
[27] http://historyofwork.iisg.nl
[28] https://wordnet.princeton.edu
[29] http://seco.cs.aalto.fi/projects/severi

was started in the EU COST project *Reassembling the Republic of Letters*[30]. Thanks to Vanhat Norssit for funding the digitization of the register and opening the data.

References

1. ter Braake, S., Fokkens, A., Sluijter, R., Declerck, T., Wandl-Vogt, E. (eds.): BD2015 Biographical Data in a Digital World 2015. CEUR Workshop Proceedings (2015), http://ceur-ws.org/Vol-1272/
2. van de Camp, M.M.: A Link to the Past. Ph.D. thesis, Tilburg University (2016), http://www.taalmonsters.nl/pdf/phd-thesis.pdf
3. Doerr, M.: The CIDOC CRM – an ontological approach to semantic interoperability of metadata. AI Magazine 24(3), 75–92 (2003)
4. van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G.: Design and use of the simple event model (SEM). Web Semantics: Science, Services and Agents on the World Wide Web 9(2), 128–136 (2011)
5. Heino, E., Tamper, M., Mäkelä, E., Leskinen, P., Ikkala, E., Tuominen, J., Koho, M., Hyvönen, E.: Named entity linking in a complex domain: Case second world war history. In: Language, Technology and Knowledge, June 19–20. pp. 120–133. Springer-Verlag (2017)
6. Hyvönen, E., Alonen, M., Ikkala, E., Mäkelä, E.: Life stories as event-based linked data: Case semantic national biography. In: Proceeedings of ISWC 2014 Posters & Demonstrations Track. pp. 1–4. CEUR Workshop Proceedings (2014), http://ceur-ws.org/Vol-1272/
7. Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J., Mäkelä, E.: WarSampo Data Service and Semantic Portal for Publishing Linked Open Data about the Second World War History. In: The Semantic Web – Latest Advances and New Domains (ESWC 2016). pp. 758–773. Springer-Verlag (2016)
8. Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., Sirola, L.: Reassembling and enriching the life stories in printed biographical registers: Norssi high school alumni on120-133 the semantic web. In: Language, Technology and Knowledge, June 19–20. pp. 113–119. Springer-Verlag (2017)
9. Szekely, P., Knoblock, C.A., Yang, F., Zhu, X., Fink, E.E., Allen, R., Goodlander, G.: Connecting the Smithsonian American Art Museum to the Linked Data Cloud. In: Extended Semantic Web Conference. pp. 593–607. Springer-Verlag (2013)
10. Van Leeuwen, M.H., Maas, I., Miles, A.: Creating a historical international standard classification of occupations an exercise in multinational interdisciplinary cooperation. Historical Methods: A Journal of Quantitative and Interdisciplinary History 37(4), 186–197 (2004)
11. Verboven, K., Carlier, M., Dumolyn, J.: A short manual to the art of prosopography. In: Prosopography Approaches and Applications. A Handbook, pp. 35–70. Unit for Prosopographical Research (Linacre College) (2007), http://dx.doi.org/1854/8212

[30] http://www.republicofletters.net

Publication III

Petri Leskinen, Eero Hyvönen, and Jouni Tuominen. Analyzing and Visualizing Prosopographical Linked Data Based on Short Biographies. In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)*, Antske Fokkens, Serge ter Braake, Ronald Sluijter, Paul Arthur, Eveline Wandl-Vogt (editors), pages 39–44, CEUR Workshop Proceedings, Linz, Austria, June 2018, online http://ceur-ws.org/Vol-2119/paper7.pdf.

Analyzing and Visualizing Prosopographical Linked Data Based on Biographies

Petri Leskinen[1], **Eero Hyvönen**[1,2], **and Jouni Tuominen**[1,2]

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

Abstract

This paper shows how faceted search on biographical data can be utilized as a flexible basis for filtering target groups of people and, in particular, how generic data analysis and visualization tools can then be applied for solving prosopographical research questions based on the filtered data. This idea is demonstrated and evaluated in practice by presenting two application case studies: 1) linked data extracted from a printed registry of over 10 000 alumni (1867–1992) of the prominent Finnish high school Norssi, and 2) a knowledge graph extracted from 13 000 short biographies of significant Finnish people (from 3rd century to present times) in the National Biography of Finland. In both cases, the data is enriched by linking their entities with several other external datasets.

Keywords: Linked Data, Data Visualization, Biography, Prosopography

1. Prosopographical Method

Biographies describe life stories of particular people of significance, with the aim of getting a better understanding of their personality and actions, e.g., to understand their motives (Roberts, 2002). In contrast, the focus of *prosopography* is to study life histories of groups of people in order to find out some kind of commonness or average in them (Verboven et al., 2007). For example, the research question may be to find out what happened to the students of a school before the World War II in terms of social ranking, employment, or military involvement after their graduation.

The prosopographical research method (Verboven et al., 2007, p. 47) consists of two major steps. First, a target group of people is selected that share desired characteristics for solving the research question at hand. Second, the target group is analyzed, and possibly compared with other groups, in order to solve the research question.

In our earlier paper (Hyvönen et al., 2017) we presented an application case study where data from a printed collection of over 10,000 short biographies (registry entries) of Norssi high school alumni were extracted and transformed into Linked Open Data, enriched by data linking to 10 external data sources, and published in a SPARQL[1] endpoint. A semantic faceted search engine and browser was developed for searching and filtering people and biographies that were enriched with internal and external linking for biographical research. Application of the same idea to the dataset of the Semantic National Biography of Finland (2014–2017) was considered in(Hyvönen et al., 2018), and the underlying data model was presented in Leskinen et al. (2017).

This paper extends this line of research by showing how the filtered target group of faceted search can be utilized as a basis for prosopographical research using different kind of data-analytic tools for solving prosopographical research questions. Such tools may involve, e.g., methods of network analysis (Easley and Kleinberg, 2010; Hanneman and Riddle, 2005) and visualizations (Dadzie and Rowe, 2011; Kehrer and Hauser, 2013).

The main contribution of this paper is to test and demonstrate the prosopographical method in practice by presenting how various data visualization tools using Google Charts and Google Maps can be integrated with the SPARQL endpoint allowing the end user to filter out target groups of people and biographies, and then to study them. In addition to providing statistical analyses of person groups, an interesting use case identified here is to compare analyses and visualizations based on different subgroups, e.g., people with same profession during different eras.

The paper is organized as follows. First, prosopographical analyses and visualizations are presented and discussed for the two linked datasets and applications using the approach outlined above: the Norssi high school alumni on the Semantic Web and the Semantic National Biography of Finland. After this contributions of the work in relation to related research are summarized and directions for further research are outlined.

2. Norssi Alumni Application

The Norssi alumni data service is available as linked open data at the Linked Data Finland platform[2], including some 892,000 triples about 131,000 resources. The digitization, "lodification", and the Vanhat Norssit Portal[3] is described in more detail in Hyvönen et al. (2017). The datasets consist of 10 137 person records, enriched with graphs of relating career events and family relations, and vocabularies of titles, schools, companies, medals, and hobbies. These additional data were extracted automatically from the short biographical descriptions of a printed book using OCR and text extraction and cleaning tools based on regular expressions.

The ontology model representing people and their biographical information in the Norssit alumni knowledge

[1]SPARQL Protocol and RDF Query Language, https://www.w3.org/TR/sparql11-query/

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

[3]http://www.norssit.fi/semweb

graph is based on the Bio CRM data model[4] (Tuominen et al., 2018), which has been developed to facilitate and harmonize the representation of biographies and cultural heritage data on the Semantic Web. Bio CRM is a domain specific extension of CIDOC CRM[5] (Doerr, 2003), the event-based ISO standard for representing and harmonizing Cultural Heritage data. It includes structures for basic data of people, personal relations, professions, and events with participants in different qualified roles. Bio CRM makes a distinction between enduring unary roles of actors, their enduring binary relationships, and perduring events, where the participants can take different roles modeled as a role concept hierarchy. The ontology and data infrastructure used for the Norssi dataset are described in detail in Leskinen et al. (2017).

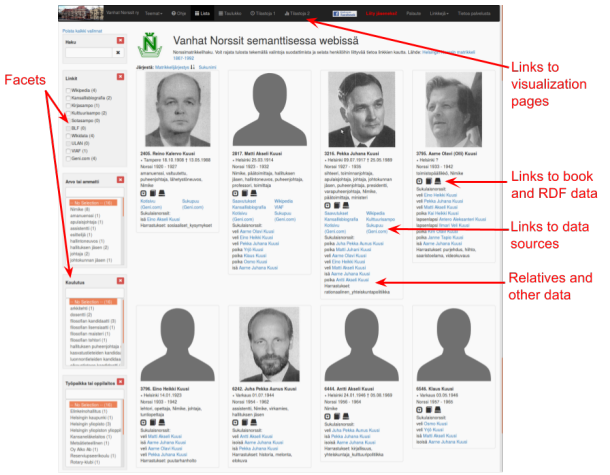


Figure 1: Faceted search for short biographies in the alumni register Norssit 1867–1992.

The Vanhat Norssit Portal contains two search interfaces, person pages, and two pages for statistical visualizations. The search interface (Fig. 1) is based on SPARQL Faceter (Koho et al., 2016), a tool for creating faceted search interfaces on a SPARQL endpoint. The interface allows the user to filter the results based on, e.g., people's education, profession, place of birth, or on which external databases he or she has been linked to.

For analyzing and visualizing data statistics of a filtered target group of people, we created two views based on Google Chart[6] diagrams. On the first visualization page[7], the popularity of the most common educations (Fig. 2), universities and colleges, professions, and employers after the graduation of the alumni are shown as four pie charts. By making filtering selections on the facets, the graphics are updated accordingly. For example, by selecting "professor" on the profession facet the employers of the 258 professors in the data can be seen on the employer pie chart. On the same page, there is also a Sankey diagram depicted in Fig. 3 that shows a list of universities on the left side and the corresponding educational titles (e.g., MSc in Technology, Doctor of Medicine, etc.) on the right. From this visualization one can see which titles were obtained from which universities regarding the filtered target group. The highlighted path in Fig. 3 shows, e.g., the connection from the University of Helsinki to Bachelor of Arts when no filtering choices have been made.

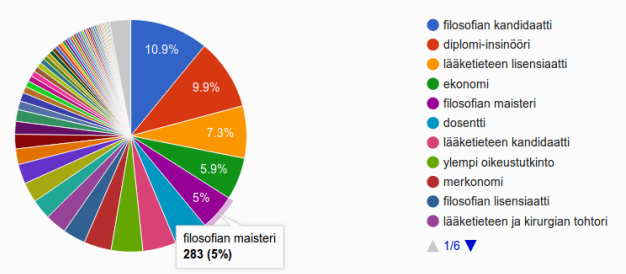


Figure 2: Pie chart showing the most common educations among high school alumni.

On the second visualization page[8], there are first two histograms showing years of enrollment and matriculation of the target group. Below these, three multi-column charts show the most popular universities and colleges, employers, and occupations of the filtered people on a decade by decade basis. For example, from the histogram representing the years of enrollment (Fig. 4) one can see that when education in Norssi was started, a lot of pupils from other schools moved to Norssi (first high bar on the left). Also the changes made in the Finnish school system in the 1970's are clearly visible as very low enrollment rates. Fig. 5 depicts the most popular employers. It shows a great and interesting variation of companies and organizations at different times: in the late 1800's the Finnish State Railways (Valtion Rautatiet, blue columns) was the most popular employer, but declined soon probably because the main railway connections in Finland were built in 1850–1900.[9] The Finnish Defense Forces (Puolustusvoimat, green columns), on the other hand, has its highest peek during the Second World War. After this the banking industry and the city of Helsinki became major employers for Norssi alumni.

The facet for links to external datasets provides also an interesting option for selecting target groups. For example, a student in the school may ask herself/himself the question: where should I work if I want to become famous and get an entry in the National Biography? By making the selection "National Biography" on the facet and then looking at the employer multi-column chart one can get an idea of where to work in order to be included in the National Biography.

The official motto of the Norssi high school is *Non scholae sed vitae* (not for school, but for life). Data analytics based on the linked data service now provides new insights on what actually happened to the school alumni in life after graduation in a prosopographical sense.

[4] http://seco.cs.aalto.fi/projects/biographies/
[5] http://cidoc-crm.org
[6] https://developers.google.com/chart/
[7] http://www.norssit.fi/semweb/#!/visualisointi
[8] http://www.norssit.fi/semweb/#!/visualisointi2
[9] https://en.wikipedia.org/wiki/History_of_rail_transport_in_Finland

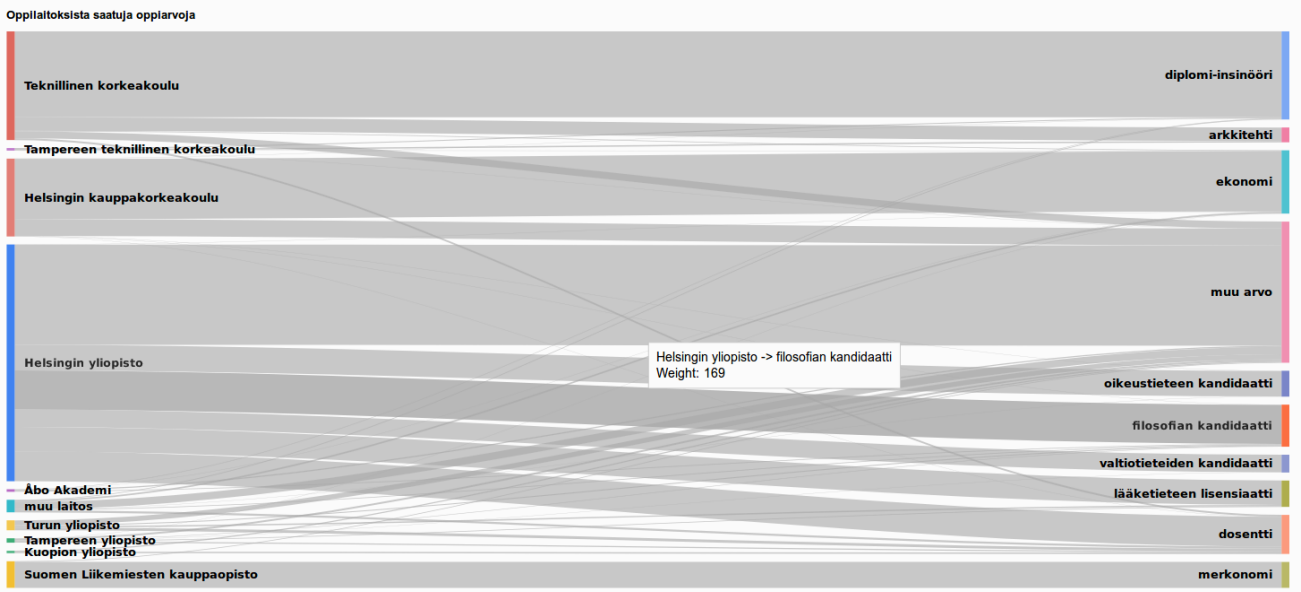


Figure 3: Sankey diagram showing the linkage between the university and the education.

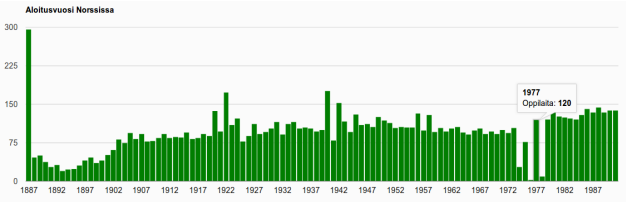


Figure 4: Column chart showing the amount of pupils by enrollment year.

3. Semantic National Biography of Finland

The National Biography of Finland[10] consists of biographies of notable Finnish people throughout history (200–2018). The biographies describe the lives and achievements of these historical and contemporary figures, containing vast amounts of references to notable Finnish and foreign figures, including internal links to other biographies of the National Biography of Finland. In addition, the text contains references to historical events, notable works (such as paintings, books, music, and acting), places (such as place of birth and death), organizations, and dates.

In this case, the texts and data were available in a database in a semi-structured form. As in the Norssi case above, the texts were transformed into RDF form by extracting entities from the semi-structured texts, and the result was uploaded into a SPARQL endpoint of the Linked Data Finland service.

The underlying ontology model represents people and their biographical information. A natural choice for modeling life stories is the event-based approach where a person's life is seen as a sequence of spatio-temporal, possibly interlinked events from birth to death (and beyond). The events are modeled according to the Bio CRM model (Tuominen et al., 2018), and the person ontology is compatible with the Getty ULAN LOD[11] model.

The source data consists (at the moment) of fields extracted from the original database dump in CSV format. In the simplest cases, the value of a data field is directly indicated by the value of a property, e.g., date or place of birth. However, most of the structured knowledge was extracted from short snippets of text in the end of each biography describing major life events of the protagonist, such as graduation from a university, designing a building, publishing a book, getting a honorary medal, etc. The resulting knowledge graph includes 13 144 people with a biographical description in the National Biography, 51 243 relating people mentioned in the biographies, and 977 authors of the biographies. At the moment, the data includes 37 730 births, 25 552 deaths, and 102 300 other biographical events. In addition to that there are 51 937 family relations, 4953 places, 3101 occupational titles, and 2938 companies extracted from the source data. (Hyvönen et al., 2018) On top of the data service, a search interface (Fig. 6) using the SPARQL Faceter tool (Koho et al., 2016) and AngularJS[12] framework was created. It can be used for finding individual biographies and for filtering out target groups for prosopography.

For biographical research, we created for each person entry page two tabs: one for the textual description of the person with additional data links, and one for a spatio-temporal visualization of the life events of the person using a map and a timeline. For prosopography, there is 1) a page for studying the events of the target group, and 2) a page for visualizing statistics of the filtered people. The application will be opened to the public in September 2018.

Fig. 7 depicts an example of a person's map-timeline page.

[10] https://kansallisbiografia.fi/english/national-biography

[11] http://www.getty.edu/research/tools/vocabularies/lod

[12] http://angularjs.org

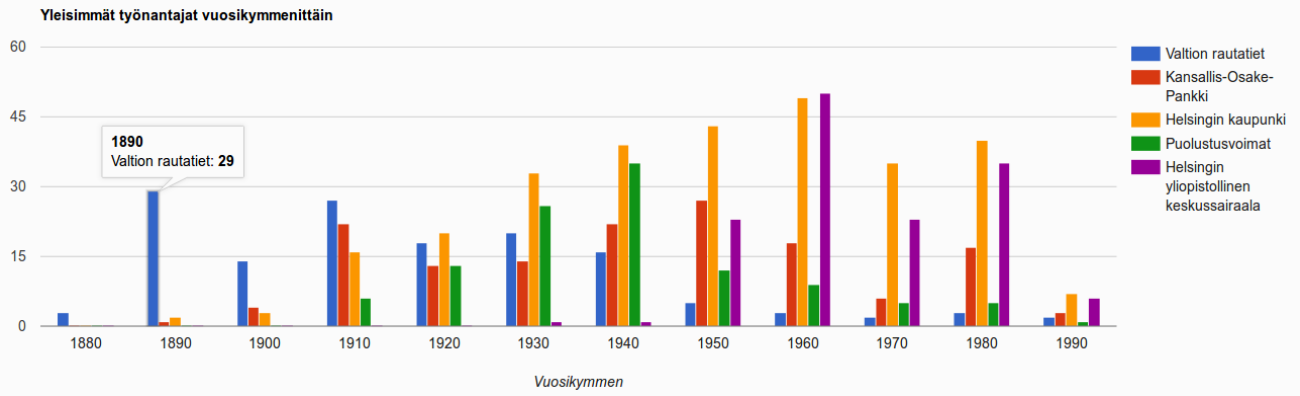


Figure 5: Column chart showing the most common employers.

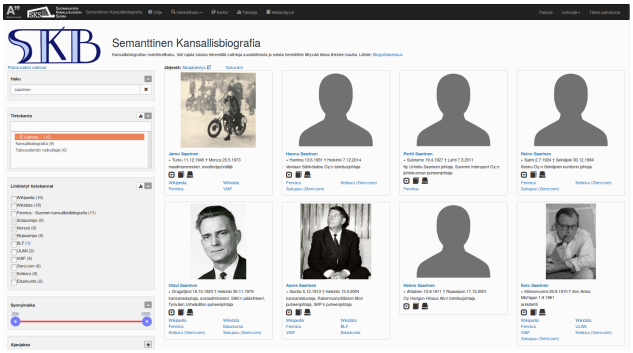


Figure 6: Main page of the Finnish National Biography.

There is a chronological list of life events on the left column. Events with known locations are shown on the map, and below there is a timeline showing the timespan of the events. The timeline spans from a person's birth to death, and shows when the career highlights have taken place. There are four horizontal lines in the timeline for separating different categories of biographical events, each represented in a different color: family events (e.g., getting married, having children), career events (e.g., education, professional experience), achievements, and mentions of honor. Corresponding markers on the map follow the same color schema.

When an event is hovered on the event list or on the timeline, the corresponding marker on the map gets highlighted. The size of the marker depends on the number of events related to that specific location, so the most important places for a person's career are emphasized. In the example case, the visualization is based on the biography of architect Eliel Saarinen, and Helsinki and Michigan (where he lived his later years) are emphasized. Data about the places in Finland was extracted from the Finnish Gazetteer of Historical Places and Maps (Hipla) databases and data service[13] (Ikkala et al., 2016; Hyvönen et al., 2016). Foreign placenames were linked using the Google Maps APIs[14]. For example, the locations of medieval universities in Europe, towns of the Hanseatic League[15], Finnish mansions, churches, and other well-known buildings were added to the place ontology using the Google services. The place ontology includes locations in different scales, such as countries, towns, villages, and in some cases even buildings with a known specified address.

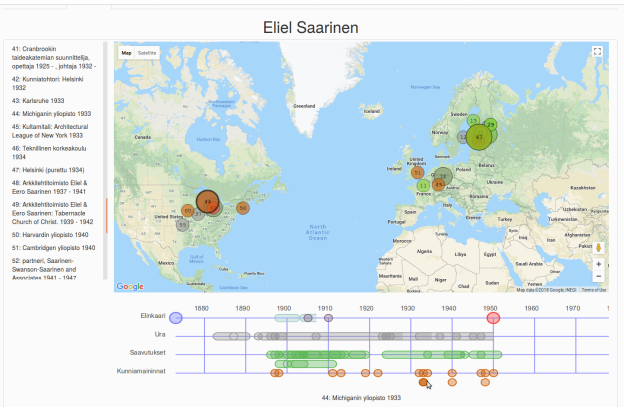


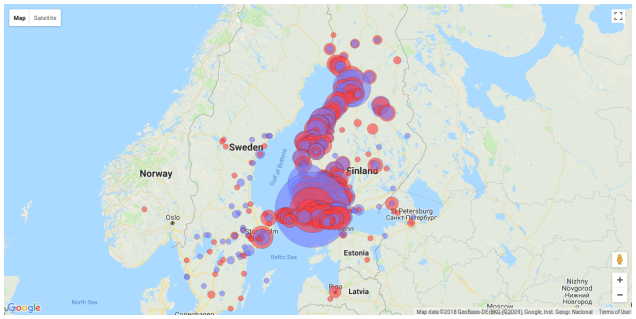
Figure 7: Map and timeline showing events related to the Finnish architect Eliel Saarinen.

As for prosopographical research, there are two different views available using Angular Google Maps[16]. The target group can be filtered by using a time span slider[17] that is included as a facet for the user to specify a desired range of years in interest. Other filtering facets include choosing person's profession, gender, dataset, related companies, related place, and linkage to external databases.

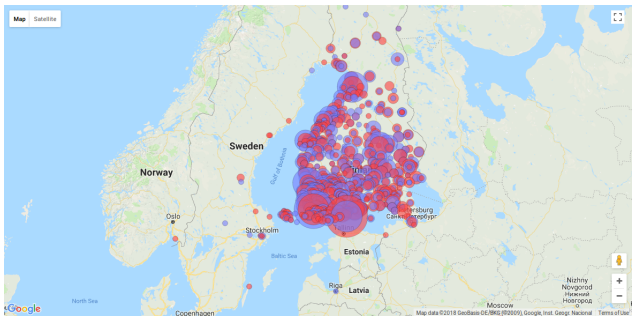
The visualizations depicted in Fig. 8, show the results of a SPARQL query corresponding to the facet selections on Angular Google Maps. The markers on the map show places of birth in blue and places of death in red color. The size of the marker corresponds to the number of events that has taken place in that particular location. Clicking on a marker opens a modal window containing a list of people who were born or died at the location.

[13] http://hipla.fi
[14] http://developers.google.com/maps/
[15] https://www.britannica.com/topic/Hanseatic-League
[16] http://angular-ui.github.io/angular-google-maps/
[17] https://github.com/angular-slider/angularjs-slider

The first selection (Fig. 8a) shows the places of birth and death of Finnish clergy 1554–1721. According to the resulting rendering, the most active areas locate along the coastal Finland with main focus on the town of Turku, which during that era was the capital of Finland, and some are scattered around Sweden. The second selection (Fig. 8b) shows the data of Finnish clergy in 1800–1920. The data does not clearly concentrate on the largest towns of Helsinki and Turku, but seem to scatter evenly around Southern Finland. During that era Finland was a part of the Russian Empire but there are only a few markers on the Russian side except at the city of St. Petersburg.



(a) The places of birth and death of Finnish clergy 1554–1721.



(b) The places of birth and death of Finnish clergy 1800–1920.

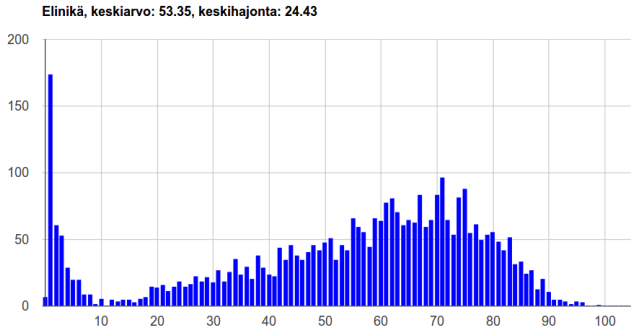
Figure 8: Two different views on the map application.

The Semantic National Biography demonstrator also includes a visualization page showing statistics as in the Norssit alumni case. The column charts in this case show (at the moment) five demographic histograms (with the mean value and standard deviation) of the target group: distribution of ages among the group, ages of marriage, ages of having the first child, the number of children, and the number of spouses.

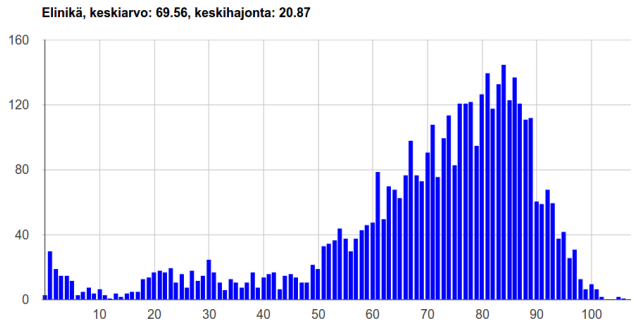
Two examples of histograms are shown in Fig. 9. The upper (a) one shows the lifespan of people who lived in 18th century, and the lower one (b) people living in 1900–1950. The two figures can be compared, e.g., how the amount of deaths among young children has decreased and how the average age has increased between the two time periods.

4. Discussion, Related Work, and Future Research

This paper demonstrated how Linked Data can be used as a basis for representing biographical registries and for filtering out target groups of persons of interest. Our particular goal was to show by a series of examples, how a SPARQL



(a) Lifespan of people lived in 1700–1800.



(b) Lifespan of people lived in 1900–1950.

Figure 9: Two different views of statistical visualizations.

endpoint can be used for data analysis and visualizations in biographical and prosopographical research. According to our practical experiences, the technology is very useful and handy to use for this after learning the basics of Linked Data standard publishing principles.

Previous works of applying Linked Data technologies to biographical data include, e.g., Larson (2010), Biographynet.nl[18] (Ockeloen et al., 2013), and our own earlier work (Hyvönen et al., 2014). The conference proceedings (ter Braake et al., 2015) include several papers on bringing biographical data online, on analyzing biographies with computational methods, on group portraits and networks, and on visualizations. Applying Linked Data principles to cultural heritage data (Hyvönen, 2012) and historical research (Meroño-Peñuela et al., 2015) has been a promising approach to solve the problems of isolated and semantically heterogeneous data sources. Also a number of previous research exists in Linked Data visualization (Bikakis and Sellis, 2016; Dadzie and Rowe, 2011).

An important component in representing biographical data is representing people and their networks, so the next part of our work is applying the methods of computational network analyses on the data. Representing biographies as linked data provides several approaches for creating such networks. For example, the biographical texts can be analyzed and people mentioned in text descriptions can be used as links in the person interrelation graph.

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

Acknowledgements

The presented research is part of the Severi project[19], funded mainly by Business Finland. Developing the National Biography of Finland is also part of the Open Science and Research Programme[20], funded by the Ministry of Education and Culture of Finland.

5. References

Nikos Bikakis and Timos Sellis. 2016. Exploration and visualization in the web of big linked data: A survey of the state of the art. In *Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference*. CEUR Workshop Proceedings, Vol-1558.

Aba Sah Dadzie and Matthew Rowe. 2011. Approaches to visualising Linked Data: A survey. *Semantic Web*, 2(2):89–124.

Martin Doerr. 2003. The CIDOC CRM – an ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3):75–92.

David Easley and Jon Kleinberg. 2010. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press.

Robert A. Hanneman and Mark Riddle. 2005. *Introduction to social network methods*. University of California, Riverside, CA. http://faculty.ucr.edu/~hanneman/.

Eero Hyvönen, Miika Alonen, Esko Ikkala, and Eetu Mäkelä. 2014. Life stories as event-based linked data: Case Semantic National Biography. In *Proceedings of ISWC 2014 Posters & Demonstrations Track*. CEUR Workshop Proceedings, October.

Eero Hyvönen. 2012. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, Palo Alto, CA, USA.

Eero Hyvönen, Esko Ikkala, and Jouni Tuominen. 2016. Linked data brokering service for historical places and maps. In *Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe)*, pages 39–52. CEUR Workshop Proc. Vol 1608.

Eero Hyvönen, Petri Leskinen, Erkki Heino, Jouni Tuominen, and Laura Sirola. 2017. Reassembling and enriching the life stories in printed biographical registers: Norssi high school alumni on the Semantic Web. In *Language, Technology and Knowledge. First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017*. Springer-Verlag.

Eero Hyvönen, Petri Leskinen, Minna Tamper, Jouni Tuominen, and Kirsi Keravuori. 2018. Semantic National Biography of Finland. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018)*, pages 372–385. CEUR Workshop Proceedings, Vol-2084, March.

Esko Ikkala, Jouni Tuominen, and Eero Hyvönen. 2016. Contextualizing historical places in a gazetteer by using historical maps and linked data. In *Proceedings of Digital Humanities 2016, Krakow, Poland, short papers*, pages 573–577.

Johannes Kehrer and Helwig Hauser. 2013. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE transactions on visualization and computer graphics*, 19(3):495–513.

Mikko Koho, Erkki Heino, and Eero Hyvönen. 2016. SPARQL Faceter—Client-side Faceted Search Based on SPARQL. In Raphaël Troncy, Ruben Verborgh, Lyndon Nixon, Thomas Kurz, Kai Schlegel, and Miel Vander Sande, editors, *Joint Proc. of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop*. CEUR Workshop Proceedings, Vol-1615.

Ray Larson. 2010. Bringing lives to light: Biography in context. Final Project Report, University of Berkeley.

Petri Leskinen, Jouni Tuominen, Erkki Heino, and Eero Hyvönen. 2017. An ontology and data infrastructure for publishing and using biographical linked data. In *Proceedings of the Workshop on Humanities in the Semantic Web (WHiSe II)*, pages 15–26. CEUR Workshop Proceedings, Vol-2014.

Albert Meroño-Peñuela, Ashkan Ashkpour, Marieke Van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach, and Frank Van Harmelen. 2015. Semantic technologies for historical research: A survey. *Semantic Web*, 6(6):539–564.

Niels Ockeloen, Antske Fokkens, Serge ter Braake, Piek Vossen, Victor De Boer, Guus Schreiber, and Susan Legêne. 2013. BiographyNet: Managing provenance at multiple levels and from different perspectives. In *Proceedings of the 3rd International Conference on Linked Science (LISC'13)*, pages 59–71. CEUR Workshp Proceedings, Vol-1116.

Brian Roberts. 2002. *Biographical Research*. Understanding social research. Open University Press.

Serge ter Braake, Ronald Sluijter Anstke Fokkens, Thierry Declerck, and Eveline Wandl-Vogt, editors. 2015. *BD2015 Biographical Data in a Digital World 2015*. CEUR Workshop Proceedings, Vol-1399.

Jouni Tuominen, Eero Hyvönen, and Petri Leskinen. 2018. Bio CRM: A data model for representing biographical data for prosopographical research. In *BD2017 Biographical Data in a Digital World 2017, Proceedings*. CEUR Workshop Proceedings.

Koenraad Verboven, Myriam Carlier, and Jan Dumolyn. 2007. A short manual to the art of prosopography. In *Prosopography Approaches and Applications. A Handbook*, pages 35–70. University of Ghent.

[19] http://seco.cs.aalto.fi/projects/severi

[20] https://openscience.fi

Publication IV

Petri Leskinen, Goki Miyakita, Mikko Koho, and Eero Hyvönen. Combining Faceted Search with Data-analytic Visualizations on Top of a SPARQL Endpoint. In *Proceedings of VOILA 2018, Monterey, California. CEUR Workshop Proceedings, Vol. 2187*, Valentina Ivanova, Patrick Lambrix, Steffen Lohmann, Catia Pesquita (editors), Monterey, CA, USA, August 2018, online `https://ceur-ws.org/Vol-2187/paper5.pdf` .

Combining Faceted Search with Data-analytic Visualizations on Top of a SPARQL Endpoint

Petri Leskinen[1], Goki Miyakita[2], Mikko Koho[1], and Eero Hyvönen[1,3]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland
[2] KMD Research Institute, Keio University, Japan
[3] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
http://seco.cs.aalto.fi, http://heldig.fi

Abstract. This paper discusses practical experiences on creating data-analytic visualizations in a browser, on top of a SPARQL endpoint based on results of faceted search. Four use cases related to Digital Humanities research in proposography are discussed in which the SPARQL Faceter tool was used and extended in different ways. The Faceter tool allows the user to select a group of people with shared properties, e.g., people with the same place of birth, gender, profession, or employer. The filtered data can then be visualized, e.g., as column charts, with business graphics, sankey diagrams, or on a map. The use cases examine the potential of visualization as well as automated knowledge discovery in Digital Humanities research.

Keywords: Linked Data, Visualization, Biography, Prosopography, Knowledge Discovery

1 Client-side Faceted Search on a SPARQL Endpoint

Faceted search and browsing [1,12], known also as view-based search [10] and dynamic hierarchies [11], has become a norm in web applications. The idea here is to index data items along orthogonal category hierarchies, i.e., facets [4] (e.g., places, times, document types etc.) and use them for searching and browsing: the user selects in free order categories on facets, and the data items included in the selected categories are considered search results. After each selection, a count is computed for each category showing the number of results, if the user next makes that selection. In this way, search is guided by avoiding annoying ”no hits” results. Moreover, hit distributions on facets provide the end-user with data-analytic views on what kind of items there are in the underlying database. Faceted search is especially useful on the Semantic Web where hierarchical ontologies used for data annotation provide a natural basis for facets, and reasoning can be used for mapping heterogeneous data to facets [2]. The idea of combining faceted search and visualizations has been applied, e.g., in ePistolarium[5]. However, this application is not based on Linked Data unlike ours [3,4,8,9].

[4] The idea of facets dates back to the Colon Classification system of S. R. Ranganathan in library science, published in 1933.

[5] http://ckcc.huygens.knaw.nl/epistolarium/

Faceted search can be implemented with server-side solutions, such as Solr[6], Sphinx[7], and ElasticSearch[8], and higher level tools, such as vuFind[9]. However there is a lack of light-weight client-side faceted search tools or components that are able to search large datasets directly from a SPARQL endpoint. Such a tool is useful, because it can be used easily on virtually any open SPARQL endpoint on the web without any need for server side programming and access rights. This paper presents such a tool, SPARQL Faceter, a web component for implementing faceted search applications efficiently in a browser, based only on a standard SPARQL API. We extend our earlier short paper of the tool [6] by 1) showing in more detail how the tool is used and works, by 2) explaining novelties in its latest version, and 3) especially show how the tool and faceted search can be extended with different kind of data-analytic visualizations.

As a proof of concept, four use case studies of data visualization are discussed from a SPARQL Faceter perspective: 1) WarSampo, using cultural heritage materials of World War II in Finland [3]. 2) Norssit, on top of a Finnish high school alumni registry data [4]. 3) Semantic National Biography of Finland, based on the National biography of the Finnish literature society [8]. 4) U.S. Congress Prosopographer, utilizing biographical records of U.S. Congress legislators [9]. In these cases, the following two-step prosopographical research method [13, p. 47] is supported where the goal is to find out some kind of commonness or average in selected *target groups* of people. First, a target group of people is selected that share desired characteristics for solving the research question at hand. Second, the target group is analyzed, and possibly compared with other groups, in order to solve the research question. For finding target groups, faceted search is used, and then visualizations are created in order to analyze their characteristics.

The rest of the paper is organized as follows. First, characteristics of SPARQL Faceter are explained with a focus on showing how it is used in practice in applications. After this, extending the tool with visuliazations is in focus. In conclusion, lessons learned and directions for further research are discussed.

2 Using and Extending SPARQL Faceter

SPARQL Faceter uses AngularJS[10] as the implementation framework. The GitHub page[11] gives instructions how to install it, and how to define the application with facets of desired type in the source code. A couple of demo examples with queries to DBpedia and WarSampo databases are provided. It can be adopted to any Linked Data publication by configuring the endpoint, property paths for facets, and queries. The SPARQL Faceter is documented in detail[12].

[6] http://lucene.apache.org/solr/

[7] http://sphinxsearch.com/blog/2013/06/21/faceted-search-with-sphinx/

[8] https://www.elastic.co/

[9] https://vufind.org/

[10] https://angularjs.org/

[11] https://github.com/SemanticComputing/angular-semantic-faceted-search

[12] http://semanticcomputing.github.io/angular-semantic-faceted-search/#/api

The data used in our applications are available as linked open data at the Linked Data Finland platform[13] for automated data publishing. Through this platform, the data is available for analyzing textual data as well as for creating semantic annotations (semi-) automatically by using data curation tools, e.g., SAHA.[14]

```
PREFIX rdf:     <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:    <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd:     <http://www.w3.org/2001/XMLSchema#>
PREFIX bioc:    <http://ldf.fi/schema/bioc/>
PREFIX rdfs:    <http://www.w3.org/2000/01/rdf-schema#>
PREFIX schema:  <http://schema.org/>
PREFIX skos:    <http://www.w3.org/2004/02/skos/core#>
PREFIX skosxl:  <http://www.w3.org/2008/05/skos-xl#>
PREFIX nbf:     <http://ldf.fi/nbf/>
PREFIX gvp:     <http://vocab.getty.edu/ontology#>
PREFIX crm:     <http://www.cidoc-crm.org/cidoc-crm/>
PREFIX rdf:     <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf:    <http://xmlns.com/foaf/0.1/>
PREFIX gvp:     <http://vocab.getty.edu/ontology#>

SELECT DISTINCT (?id AS ?id__uri) ?id__name ?value WHERE {
  # Restraints set in Faceter
  { ?id a <http://ldf.fi/nbf/PersonConcept> .
    ?id <http://xmlns.com/foaf/0.1/focus>/^<http://www.cidoc-crm.org/cidoc-crm/P98_brought_into_life>/
        <http://ldf.fi/nbf/time>/<http://vocab.getty.edu/ontology#estStart> ?slider_2 .
    FILTER (1800<=year(?slider_2) && year(?slider_2)<=2018)
    }

  # Query person's age
  ?id foaf:focus/^crm:P100_was_death_of/nbf:time [ gvp:estStart ?time ; gvp:estEnd ?time2 ] ;
      foaf:focus/^crm:P98_brought_into_life/nbf:time [ gvp:estStart ?birth ; gvp:estEnd ?birth2 ] .
  BIND (xsd:integer(0.5*(year(?time)+year(?time2)-year(?birth)-year(?birth2))) AS ?value)
  # Filter out erroneous cases
  FILTER (-1<?value && ?value<120)

  # Query for person's name
  ?id skosxl:prefLabel ?id__label .
  OPTIONAL { ?id__label schema:familyName ?id__fname }
  OPTIONAL { ?id__label schema:givenName ?id__gname }
  BIND (CONCAT(COALESCE(?id__gname, "")," ",COALESCE(?id__fname, "")) AS ?id__name)

} ORDER BY ?value ?id__fname ?id__gname
```

Fig. 1. A SPARQL example for querying people's ages

Figure 1 depicts a SPARQL query for fetching the data visualized in Fig. 2. The first block in the query pattern defines the restricted target group of the Faceter application, in this case we are interested people who were born on or after the year 1800, a choice that has been made with the timespan slider. The example follows the data model of the National Biography of Finland [8], so to query for the desired resource in the data, property paths are utilized. In the next block related events of birth are searched, and the age of a person is calculated. Possible errors in the data are filtered out by accepting only values in the range of 0 to 120 years. In the third block, the person's proper name is constructed. Some of the fields are optional, because due to the data, we cannot assume all the person entries to have both the first and the family names. The query returns JSON formatted array consisting of objects containing the URI of the resource, the person name, and age. In the application the data is converted to a JavaScript array

[13] http://ldf.fi

[14] http://demo.seco.tkk.fi/saha

suitable as input for, e.g., Google Chart tools[15], or Google Maps[16]. The output in this example case, (Fig. 2) is a column chart with age on the horizontal, and the amount of people on the vertical axis. A mouse click on any of the columns shows a modal list of all people having that age.

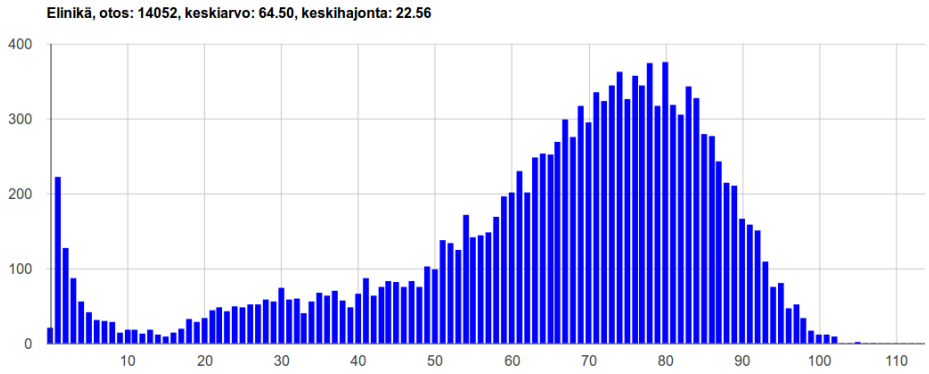


Fig. 2. Lifespan of people lived in 1800–2018

3 Applications

In this section examples of visualizations on top of the SPARQL Faceter tool in different applications are shown and discussed.
is the first semantic portal for serving and publishing large heterogeneous sets of linked open data about the World War II (WW2)[17]. To create a global view of the war, and to attain a deeper understanding of its history, the portal contains. e.g., some 95 000 death records of WW2 casualties. This in-use portal includes 8 different application perspectives through different datasets, and had 130 000 users in 2017.

Fig. 3 shows a screenshot of the faceted search application in the casualties perspective. The data is laid out in a table-like view. Facets are presented on the left of the interface with string search support. The number of hits on each facet is calculated dynamically and shown to the user, and selections leading to an empty result set are hidden. In Fig. 3, seven facets and the results are shown, where the user has selected "widow" in the marital status facet, focusing the search down to 278 killed widows that are presented in the table with links to further information.

The faceted search is used not only for searching but also as a flexible tool for researching the underlying data. In Fig. 3, the hit counts immediately show distributions

[15] https://developers.google.com/chart/
[16] https://cloud.google.com/maps-platform/
[17] https://www.sotasampo.fi/en/

of the killed widows along the facet categories. For example, the facet “Number of children” shows that one of the deceased had 10 children and most often (in 88 cases) widows had one child. If we next select the category “one child” on its facet, we can see that two of the deceased are women and 86 are men in the gender facet. In the latest version of SPARQL Faceter, each facet component has a push button for visualizing the distributions with Google pie charts.

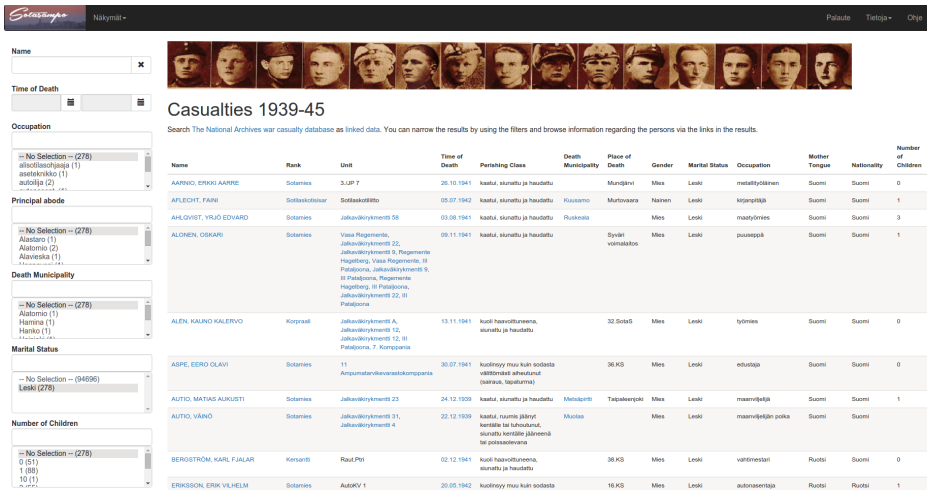


Fig. 3. The faceted search interface of death records with one selected facet. The left side contains the facets, displaying available categories and the amount of death records for each casualty. Death records matching the current facet selections are shown as a table.

2) Norssit dataset consists of a register with over 10 000 alumni of the prominent Finnish high school “Norssi” in 1867–1992. The register was transformed into RDF, was enriched by data linking, was published as a linked data service, and is provided to end users via a faceted search engine and browser for studying lives of historical persons and for prosopographical research. [4]

The Norssit portal[18] contains two pages of visualizations[19]. The pages use Google Charts showing search results as pie or column charts or sankey diagrams [7]. An example of rendering the most common employers on different decades is depicted in Fig. 4.

3) Semantic National Biography of Finland The National Biography of the Finland[20] consists of biographies of notable Finnish people throughout history. The biographies describe the lives and achievements of historical figures, containing vast amounts

[18] http://www.norssit.fi/semweb

[19] http://www.norssit.fi/semweb/#!/visualisointi,http://www.norssit.fi/semweb/#!/visualisointi2

[20] http://kansallisbiografia.fi

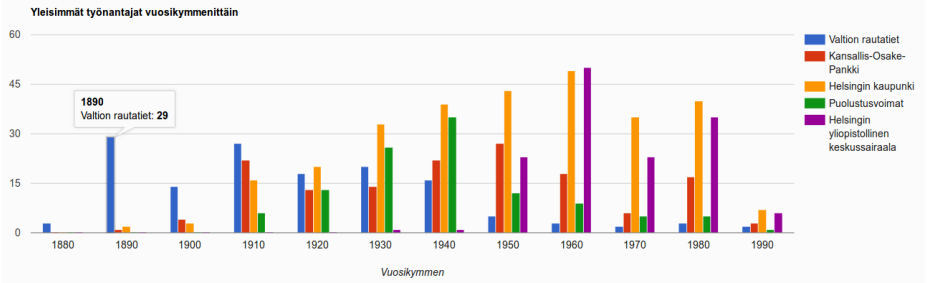


Fig. 4. Column chart showing the most common employers and their changes in time.

of references to notable Finnish and foreign figures, including internal links to other biographies. [5]

To support the prosopographical research, the portal contains pages with faceted search where the data is visualized on Google Maps, or as column charts [7]. An example of rendering the query results on Google Maps is depicted in Fig. 5. The portal also has a faceted search page for linguistic analysis of the vocabulary used in biographical descriptions.

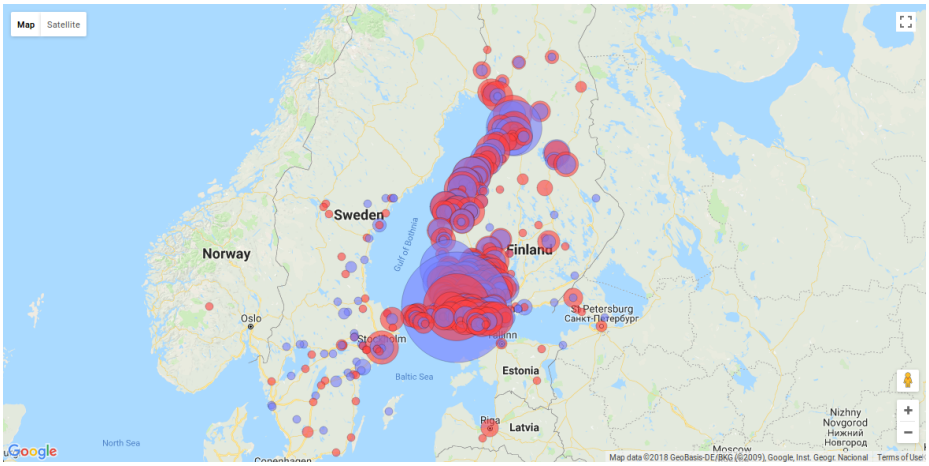


Fig. 5. Places of birth and death of 17th century Finnish clergy

4) U.S. Congress Prosopographer This interface[21] contains biographical records of 11 987 persons who served in the U.S. Congresses from the 1st (1789) to the 115th

[21] https://semanticcomputing.github.io/congress-legislators/

(2018) one—converted and extracted from open-source data[22,23]. The interface contains four integrated tools and demonstrates how historical patterns correspond to biographical information and further intertwine with politics, economics, and historical knowledge alongside the American history.

Being adapted from the previous studies above, a novelty of this interface are the comparing visualizations. As shown in Fig. 6, a different set of target groups—in this case, the two major parties, Democrats and Republicans—can be analyzed and compared with each-other. The end user is able to find and execute new insights through the independent variables, as well as the latent biographical relationship of U.S. Congress legislators through selecting, filtering, and comparing two different accounts of histories.

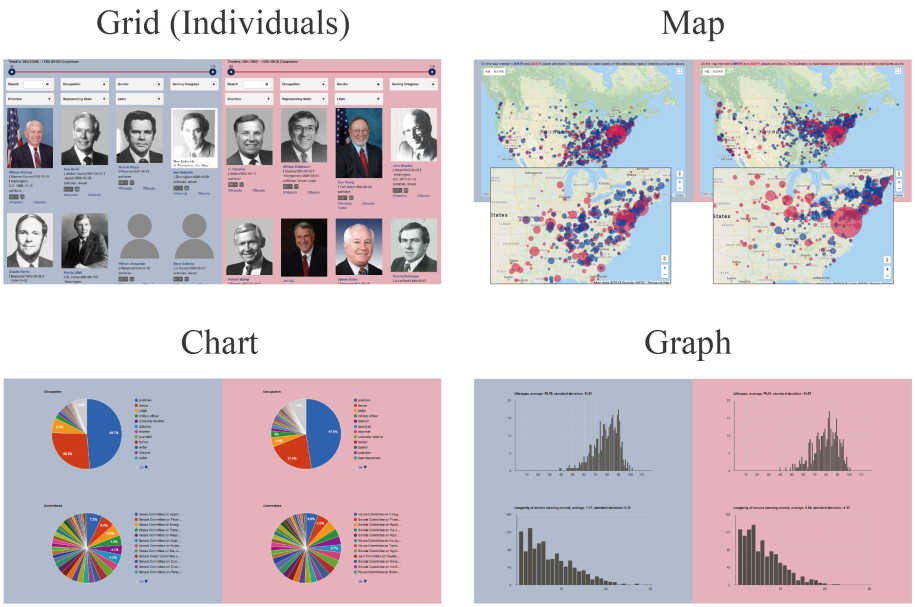


Fig. 6. Examples of Comparing Visualizations: Democratic (left) / Republican (right)

4 Discussion

Based on the applications discussed, faceted search and browsing can be combined in a useful way with various means and tools for visualization: facet selections are a

[22]https://github.com/unitedstates/congress-legislators
[23]http://k7moa.com

very flexible way to filter out result sets, and we have demonstrated that this can be done in real time using SPARQL queries in endpoints containing tens of millions of triples. Based on the query results, wrappers for data visualization tools, such as Google Charts for statistics or network analysis tools can be integrated and reused easily. By making the data analysis on the client side, computational burden can be distributed to end-user browsers, and Rich Internet Applications can be created without server-side programming. Moreover, the resulting visualizations open up ways of exploring new types of questions, and further evokes a knowledge discovery process in conducting digital humanities research.

A key challenge in this approach is how to deal with large result sets. It is usually not feasible to transfer very large result sets, say tens of thousands of casualty records in the WarSampo case, from the server to the browser. If the data is not available in the browser, it cannot of course be analyzed there. This problem is solved in SPARQL Faceter by paginating the results; the results are uploaded in pages and only when needed. The end-user should be aware about the limitation that the visualizations are based on only the data that has been uploaded. The size of the page therefore sets a limit on how large datasets can be visualized, even though very large result datasets can be queried on the server side.

The use case study WarSampo was implemented by the original SPARQL Faceter [6] while the other use cases discussed are based on its new versions with the following enhancements: 1) Every facet is now able to make its own SPARQL query (or many), which leads to better efficiency. 2) Hierarchical facets up to any number of levels are supported and more efficiently implemented. 3) Text search facet is included as a new facet type. 4) A slider facet for selecting a range of numerical values interactively can be used. 5) Facet hit distributions can be visualized using pie charts in addition to hit counts. There are also some enhancements made for visualizations. The U.S. Congress Prosopographer allows, for example, visual comparison of two groups, a functionality that should also be implemented to our ongoing project of Semantic National Biography. The different Faceter versions and extensions need to be amalgamated together in next versions of the tool and applications. Also the technical solutions for showing new types of visualization, such as social networks of people should be studied more. Still another direction for further work are the aesthetic qualities. The visualizations are generated using standardized templates, e.g., web frameworks such as Google Charts, and balancing between usability and design aesthetics needs to be studied.

Acknowledgements Thanks to Erkki Heino for implementational help regarding extending the Faceter SPARQL tool for our case studies, and to Jouni Tuominen for discussions related to the data models and data services underlying our applications. Goki Miyakita was supported by a mobility scholarship at Aalto University in the frame of the Erasmus Mundus Action 2 Project TEAM Technologies for Information and Communication Technologies, funded by the European Commission. Our research was also supported by the CSC computing services and the Severi project[24] funded mainly by Business Finland.

[24] http://seco.cs.aalto.fi/projects/severi

References

1. Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., Lee, K.P.: Finding the flow in web site search. CACM 45(9), 42–49 (2002)
2. Hyvönen, E., Saarela, S., Viljanen, K.: Application of ontology-based techniques to view-based semantic search and browsing. In: The semantic web: research and applications. First European Semantic Web Symposium (ESWS 2004). pp. 92–106. Springer-Verlag (2004)
3. Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J., Mäkelä, E.: WarSampo data service and semantic portal for publishing linked open data about the second world war history. In: The Semantic Web – Latest Advances and New Domains (ESWC 2016). pp. 758–773. Springer-Verlag (2016)
4. Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., Sirola, L.: Reassembling and enriching the life stories in printed biographical registers: Norssi high school alumni on the Semantic Web. In: Language, Technology and Knowledge. First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017. Springer-Verlag (2017)
5. Hyvönen, E., Leskinen, P., Tamper, M., Tuominen, J., Keravuori, K.: Semantic National Biography of Finland. In: Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018). pp. 372–385. CEUR Workshop Proceedings, Vol-2084 (March 2018)
6. Koho, M., Heino, E., Hyvönen, E.: SPARQL Faceter—Client-side Faceted Search Based on SPARQL. In: Troncy, R., Verborgh, R., Nixon, L., Kurz, T., Schlegel, K., Vander Sande, M. (eds.) Joint Proc. of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop. No. 1615, CEUR Workshop Proceedings (2016), http://ceur-ws.org/Vol-1615/semdevPaper5.pdf
7. Leskinen, P., Hyvönen, E., Tuominen, J.: Analyzing and visualizing prosopographical linked data based on short biographies. In: BD2017 Biographical Data in a Digital World 2017, Proceedings. CEUR Workshop Proceedings (2018)
8. Leskinen, P., Tuominen, J., Heino, E., Hyvönen, E.: An ontology and data infrastructure for publishing and using biographical linked data. In: Proceedings of the Workshop on Humanities in the Semantic Web (WHiSe II). pp. 15–26. CEUR Workshop Proceedings, Vol-2014 (2017)
9. Miyakita, G., Leskinen, P., Hyvönen, E.: U.S. Congress Prosopographer – A Tool for Prosopographical Research of Legislators (May 2018), submitted
10. Pollitt, A.S.: The key role of classification and indexing in view-based searching. Tech. rep., University of Huddersfield, UK (1998), http://www.ifla.org/IV/ifla63/63polst.pdf
11. Sacco, G.M.: Dynamic taxonomies: guided interactive diagnostic assistance. In: Wickramasinghe, N. (ed.) Encyclopedia of Healthcare Information Systems. Idea Group (2005)
12. Tunkelang, D.: Faceted search, Synthesis lectures on information concepts, retrieval, and services, vol. 1. Morgan & Claypool Publishers (2009)
13. Verboven, K., Carlier, M., Dumolyn, J.: A short manual to the art of prosopography. In: Prosopography Approaches and Applications. A Handbook, pp. 35–70. University of Ghent (2007), http://hdl.handle.net/1854/LU-376535

Publication V

Eero Hyvönen, Petri Leskinen, Minna Tamper, Heikki Rantala, Esko Ikkala, Jouni Tuominen, and Kirsi Keravuori. 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*, Pascal Hitzler, Miriam Fernández, Krzysztof Janowicz, Amrapali Zaveri, Alasdair J. G. Gray, Vanessa Lopez, Armin Haller, and Karl Hammar (editors), Lecture Notes in Computer Science, volume 11503, pages 574–589, Springer-Verlag, June 2019, online https://doi.org/10.1007/978-3-030-21348-0_37.

BiographySampo – Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research

Eero Hyvönen[1,2], Petri Leskinen[1], Minna Tamper[1], Heikki Rantala[1], Esko Ikkala[1,2], Jouni Tuominen[1,2], and Kirsi Keravuori[3]

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

Abstract. This paper argues for making a *paradigm shift* in publishing and using biographical dictionaries on the web, based on Linked Data. The idea is to provide the user with enhanced reading experience of biographies by enriching contents with data linking and reasoning. In addition, versatile tooling for 1) biographical research of individual persons as well as for 2) prosopographical research on groups of people are provided. To demonstrate and evaluate the new possibilities, we present the semantic portal "BiographySampo – Finnish Biographies on the Semantic Web". The system is based on a knowledge graph extracted automatically from a collection of 13 100 textual biographies, enriched with data linking to 16 external data sources, and by harvesting external collection data from libraries, museums, and archives. The portal was released in September 2018 for free public use at http://biografiasampo.fi.

1 National Biographical Dictionaries on the Web

Biographical dictionaries, a historical genre dating back to antiquity, are scholarly resources used by the public and by the academic community alike. Most national biographical dictionaries follow the traditional form of combining a lengthy non-structured text, often written with authorial individuality and personal insight, with a structure supplement of basic biographical facts, such as family, education, works, and so on. Biographies are an invaluable information source for researchers across the disciplines with an interest in the past. [22]

A well-known example of a biographical dictionary is the Oxford Dictionary of National Biography (ODNB)[4] with more than 60 000 lives. It was published online in 2004, and since then many biographical dictionaries have opened their editions on the Web. These include USA's American National Biography[5], Germany's Neue Deutsche

[4] http://global.oup.com/oxforddnb/info/
[5] http://www.anb.org/aboutanb.html

Biographie[6], Biography Portal of the Netherlands[7], The Dictionary of Swedish National Biography[8], and National Biography of Finland[9] [2] (NBF). In addition to biographical dictionaries of historical people there are lots of "who is who" reference books and online services focusing on describing living persons.

ODNB and other early adopters of web technology started the paradigm shift in publishing and using biographical dictionaries on the Web. This paper argues for making the next paradigm shift, i.e., to *publishing and using biographical dictionaries as Linked Data on the Semantic Web*. We present the new in-use system "BIOGRAPHYSAMPO – Finnish Biographies on the Semantic Web" based on the National Biography and other biographical databases of the Finnish Literature Society[10] interlinked with related data repositories. The idea is to 1) transform textual biographies into Linked Data (LD) by using language technology and knowledge extraction, to 2) enrich the data by linking it to internal and external data sources and by reasoning, to 3) publish the data as a LD service and a SPARQL endpoint on the web [10,13], and to 4) create end-user applications on top of the service, including data-analytic tools and visualizations for distant reading [33] of Big Data, i.e., for Digital Humanities (DH) research [9].

Today, national biography collections on the Web are used in the following traditional way: a search box or a more detailed search form is filled up specifying the person(s) whose biographies are searched for. After pushing the search button, a list of hits is shown that can be opened by clicking for close reading. BIOGRAPHYSAMPO challenges this traditional approach of publishing and using biographical dictionaries in the following ways: 1) Data from multiple biographies is provided. 2) The data is enriched by harmonizing and combining it with additional data sources, such as metadata from memory organization collections. 3) The data is enriched by reasoning for enhanced reading experience and for knowledge discovery. 4) Data-analytic and visualization tools for biographical [30] and prosopographical research [37] are provided.

In the following, the knowledge extraction process for textual bios into a harmonized knowledge graph is first described (Section 2), as well as the underlying event-based data model, datasets, and LD service (Section 3). After this, the system is considered from the end users's perspective by presenting seven application views included in the portal (Section 4). In conclusion (Section 5), the proposed paradigm change is analyzed from a Digital Humanities research perspective and related works are discussed.

2 Creating the Knowledge Graph

Knowledge Extraction The biographies in dictionaries often have two sections: the beginning is written in terms of normal full sentences, and in the end there is a concise, semi-formal summary, explicating the major events, achievements, and other biographical data about the biographee [39]. Here, for example, listings and abbreviations without verbs are widely used for explaining family relations, educational degrees, professions,

[6] http://www.ndb.badw-muenchen.de/ndb_aufgaben_e.htm
[7] http://www.biografischportaal.nl/en
[8] https://sok.riksarkivet.se/Sbl/Start.aspx?lang=en
[9] http://biografiakeskus.fi
[10] https://www.finlit.fi/en

and honorary medals.[11] An example of the semi-formal descriptions for the architect *Eliel Saarinen* is given below:

> Gottlieb Eliel Saarinen S 20.8.1873 Rantasalmi, K 1.7.1950 Bloomfield Hills, Michigan, Yhdysvallat. V rovasti Juho Saarinen ja Selma Maria Broms. P1 1898 - 1902 (ero) Mathilda Tony Charlotta Gylden (sittemmin Gesellius) S 1877, K 1921, P1 V agronomi Axel Gylden ja Antonia Sofia Hausen; P2 1904 - kuvanveistäjä Minna Carolina Louise (Loja) Gesellius S 1879, ...
> URA. Arkkitehtitoimisto Gesellius, Lindgren & Saarinen, perustajajäsen, osakas 1896–1907; Arkkitehtitoimisto Eliel Saarinen, johtaja 1907–1923; ...
> TEOKSET. Arkkitehtitoimisto Gesellius, Lindgren, Saarinen: Tallbergin talo. 1896–1898, Luotsikatu 1, Helsinki; Pariisin maailmannäyttelyn 1900 paviljonki. 1898–1900, Pariisi;

The semi-formal expressions here have uniformity in structure that can be used effectively for pattern-based information extraction: First, the person's given and family names are mentioned and after that the fields of birth and death information are separated with *S* for birth, and *K* for death. These fields contain the time and place of the event. A field beginning with *V* contains the information about the person's parents with the father followed by the mother, their names, occupations, and possible places and times of birth and death. Likewise, fields beginning with *P*, or if several *P1*, *P2* etc., carry the information of possible spouses indicating the year of marriage, and the spouse's years of birth and death. The data field may also contain information about the parents of the spouse in *PV*, *PV1*, *PV2*, etc. fields. In addition to family relations, there are descriptions of person's life time events also in a semi-formal format. The paragraph begins with a label telling if the listed events deal with his education, career (*URA*), or achievements (*TEOKSET*). The events listed are separated with a semicolon, and each event has a textual description ending with time period and place. For knowledge extraction of the semi-formal part, rules based of regular expressions were used in BIOGRAPHYSAMPO.

The pipeline for the free text part was built using pre-existing NLP tools [34]. The process consists of linguistic analyses (such as tokenization and morphological tagging) and converting the document structures and the linguistic data into RDF. The NLP Interchange Format (NIF)[12] [11] supplements the RDF representation with a Core Ontology that provides classes and properties to describe the relations between texts and documents. This provides flexibility and structure to divide a document into paragraphs, titles, sentences, and words that can be complemented with structural metadata supplied by NIF and linguistic information, such as lemmas and part-of-speech (POS) tags from NLP tools. In addition to the NIF format, the commonly used CIDOC CRM ISO standard, Dublin Core Metadata[13], and a custom namespace are used to supply classes and properties for describing document metadata.

BIOGRAPHYSAMPO automatically creates a narrative life story for each of the 13 100 protagonists in the biographies. [34] This story is then enriched in the following ways: 1) Links to other external biographies of the person are created for additional information and for using the linkage as a criterion for faceted search and determining target groups in prosopography. 2) The data is enriched from additional external sources, such as collection data from museums, libraries, and archives. For example, if there is

[11] In person registries [18], the whole entry text may be semi-formal.
[12] http://persistence.uni-leipzig.org/nlp2rdf/specification/core.html accessed: 13 August 2018
[13] http://dublincore.org/documents/dcmi-terms/ accessed: 13 August 2018

a painting by an artist in a collection, the corresponding artistic creation event can be added as an entry in the biographical timeline of the protagonist. 3) The data is enriched by reasoning. For example, links to persons with similar life stories are determined for recommendation links, new family relations and egocentric networks between persons are explicated, and serendipitous relations between entities such as persons and places are discovered.

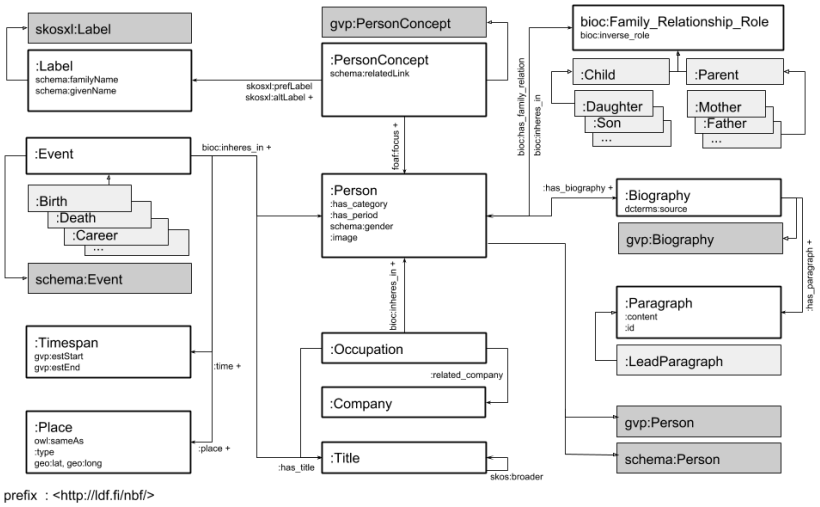


Fig. 1: Data model for BIOGRAPHYSAMPO. In addition to using *owl* and *skos(xl)* standards, namespace *bioc* refers to Bio CRM, *schema* to schema.org, *geo* to W3C Basic Geo, and *gvp* to Getty vocabulary. [36]

3 Data Model, Datasets, and Data Service

The data model used is depicted in Fig. 1. The central class in the middle is :Person. The life of a person instance is described essentially in terms of different kind of events (s)he participated in different roles in time and place (on the left side of the Figure). For the human reader, links to biographical texts are provided (on the right). However, based on the machine understandable RDF data, the reading experience can be enhanced by providing the end user with additional information related to the biographies and with tools for analyzing the lives and biographies as texts in different way, as will be shown in the following sections. The data model used is an extension of CIDOC CRM [5,26] that we call Bio CRM; see [36] for more details.

The core data includes the biography collections listed in Table 1, edited and maintained by the Biographical Centre of the Finnish Literature Society (SKS). These biographies have been written by 977 scholars from different fields. The largest collection,

the National Biography of Finland (Suomen kansallisbiografia), was first published online in 1997 and later on as a 9500 page book series of ten volumes and a separate index [2]. All biographies in Table 1 are today available via a national web service[14].

The core datasets are linked not only internally but also enriched with links to the external data sources of biographies listed in Table 2 according to the Linked Data 5-star model[15]. The links were created by comparing names and birth years and were included in our data service for additional information of persons. In addition, the data sources are used as a search facet for filtering out persons described in different data sources. In comparison to our earlier prototype [19], two new datasets were linked in the system: 1) The national bibliography Fennica[16], published by the National Library of Finland, containing the largest collection of bibliographical entries in Finland. 2) The University of Helsinki student register (1853–1899)[17].

Also data from the following datasets was harvested and partly included: 1) The open art collection data of the National Gallery of Finland[18]. 2) National bibliography of Finland Fennica. 3) Critical Edition of J.V. Snellman's works [1], published online[19] by Edita Ltd. J. V. Snellman (1806–1881) was a most prominent figure of the Finnish history in the 19th century. The data was converted into RDF and contains, e.g., some 3000 works and references to thousands of historical persons. 4) Booksampo semantic portal[20], containing linked data about virtually all Finnish fiction literature. 5) The Finnish historical ontology HISTO[21], containing linked data about important events of Finnish history. The idea here was to investigate and to show, how biographical data can be enriched by different kinds of collection contents from museums, libraries, and archives. This kind of data was instrumental, e.g., in creating the relational search application perspective of the portal [20] (to be presented in more detail later on).

Dataset name	**# of People**
National Biography of Finland	6478
Business Leaders	2235
Finnish Generals and Admirals 1809–1917	481
Finnish Clergy 1554–1721	2716
Finnish Clergy 1800–1920	1234
Sum	13144

Table 1: The biography datasets provided by the Finnish Literature Society. The biographies of the National Biography and the Finnish Clergy datasets contain semi-formal summaries.

[14] http://kansallisbiografia.fi
[15] http://5stardata.info/en/
[16] https://www.kansalliskirjasto.fi/en/news/finnish-national-bibliography-released-as-open-data
[17] https://ylioppilasmatrikkeli.helsinki.fi/1853-1899/
[18] https://www.kansallisgalleria.fi/en/avoin-data/
[19] http://snellman.kootutteokset.fi/
[20] http://kirjasampo.fi
[21] https://seco.cs.aalto.fi/ontologies/histo/

Data Source	# of Links	Description
Wikipedia	6316	http://fi.wikipedia.org
Wikidata	6505	http://www.wikidata.org
Fennica	4007	National Bibliography of Finland
BLF	1084	Biografiskt Lexikon för Finland
BookSampo	715	Finnish fiction literature LD service
WarSampo	288	Second World War LOD service and portal
ULAN	213	Union List of Artist Names Online
VIAF	2475	Virtual International Authority Files
Geni.com	5320	Family research and family tree data
Home pages	43	Personal web sites
Parliament of Finland	631	Members of Parliament of Finland 1917–2018
University of Helsinki (UH) Registry	379	Students and faculty of UH in 1853–1899
Sum	28197	

Table 2: External data sources (person pages) linked to the BIOGRAPHYSAMPO.

BIOGRAPHYSAMPO Data Service serves 13 144 biographies from which some 125 000 events, 51 937 family relations, 4953 places, 3101 professions, and 2938 companies were identified and extracted. There are also over 26 000 links to the 16 linked external biographical datasets and services, and tens of thousands of relations extracted from external sources. The biographical data contains ca. 10 million triples, and there is a separate graph of over 100 million triples representing the texts linguistically.

In order to evaluate the knowledge extraction pipeline (cf. Section 2), a test set of 135 events was manually checked with promising results: 99% of the generated data were actual events of a person's life, and 98% of events had a correct time period. We filtered out the snippets having a timespan outside of person's living years. The text snippets were also linked to our place ontology with a precision of 98%, and a recall of 77%. The process produced false positives in cases, e.g., when a company has the same name as a place. In some cases, lemmatizing a place name caused a wrong basic form, and the event did not get linked to the correct place.

The data is provided using the "7-star" Linked Data Finland platform[22] [16]. The service is based on Fuseki[23] with a Varnish Cache[24] front end for resolving URIs and serving LD in different ways. A larger vision behind our work is that by publishing openly shared ontologies and data about historical persons for everybody to use, future interoperability problems can be prevented before they arise [12]. At the moment, all data has been opened for the public to read freely. Negotiations for opening the data service as well are underway.

The data service can be used as a basis for Rich Internet Applications (RIA). A demonstration of this is the BIOGRAPHYSAMPO Portal, where *all* functionality is implemented on the client side using JavaScript, only data is fetched from the server side SPARQL endpoints. In the next section, new ways of using the biographical linked data in the portal are presented from the end-user's point of view.

[22] See http://www.ldf.fi for more details.

[23] http://jena.apache.org/documentation/serving_data/

[24] https://www.varnish-cache.org

4 New Ways for Studying Biographies

The BIOGRAPHYSAMPO Portal is not just one application, but a collection of thematic interlinked *application perspectives* to the underlying data. Different perspectives are needed [15,28] in order to address different end-user information needs properly. This idea is in contrast with large monolithic portals that may show only one view or search perspective of the data.

The portal includes seven perspectives that can be selected in the front page of the system or at any situation in the menu bar: 1) *Persons*. Faceted search view for filtering and finding biographies. 2) *Places*. Searching biographical events projected on interactive maps. 3) *Life maps*. Life events and trajectories from birth to death of person groups visualized on maps. 4) *Statistics*. Various histogram and pie chart statistics of filtered person groups. 5) *Networks*. Analyzing networks of person groups. 6) *Relations*. Finding serendipitous connections between persons and places with natural language explanations. 7) *Language*. Tools for analyzing the language used in biographies.

Many perspectives of the portal support the prosopographical research method [37, p. 47] that consists of two major steps. First, a target group of people that share desired characteristics is selected for solving the research question at hand. Second, the target group is analyzed, and possibly compared with other groups, in order to solve the research question. To support prosopography, BIOGRAPHYSAMPO employs faceted search for filtering out target groups. Once the group has been determined, various generic data-analytic tools and visualizations can be applied to it. In below, the major functionalities of the portal's perspectives are explained from the end user's view point.

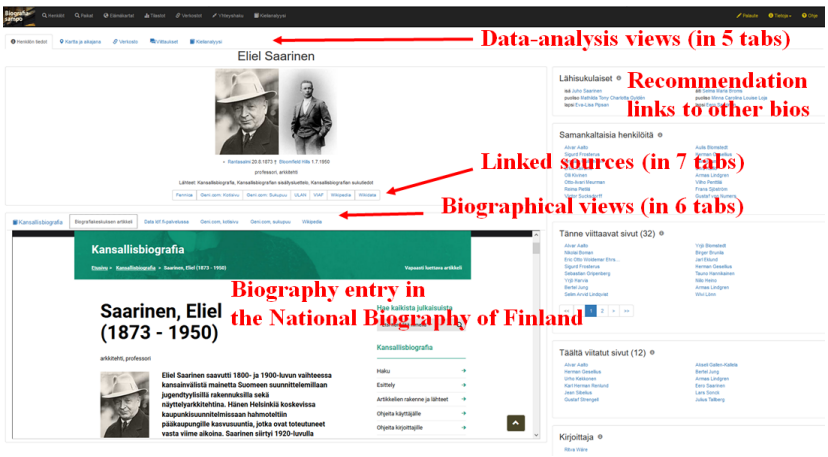


Fig. 2: Home page of Eliel Saarinen (1873–1950).

1. Persons The basic use case in biography collections is to find a person's biography to be read. In addition to supporting traditional name string based search, the

Persons view features a full-blown faceted search engine on top of a SPARQL endpoint. Here properties, such as profession, place of birth, place of education, working organization, and other criteria can be used for filtering down persons of potential interest. After each facet category selection, the hit counts on all facets are calculated, so that the user never ends up in a ”no hits” situation. Furthermore, the hit counts on the facets provide useful statistical information about the distributions of biographies along the orthogonal facets. The distributions can also be visualized as interactive pie charts by a click on a special symbol. The faceted search engine was implemented by developing a new version of the Faceter tool [23].

BIOGRAPHYSAMPO generates for each person in the system a global ”home page” for enhanced reading experience by enriching data from various interlinked data sources and by reasoning. After finding a person of interest, BIOGRAPHYSAMPO provides the user with an enriched reading view of his or her life based on 1) data linking and 2) reasoning. Fig. 2 shows as an example the home page of Eliel Saarinen (1873–1950), a prominent Finnish architect. The page contains six tabs providing different biographical views of the person, here two pages based on the NBF, data at the Linked Data Finland service, a genealogical family tree and home page by the Geni.com service, and the Finnish Wikipedia article. The entry is linked to seven external data sources on the web. On the right, recommendation links to related biographies are given, e.g., to similar biographies based on their linguistic content.

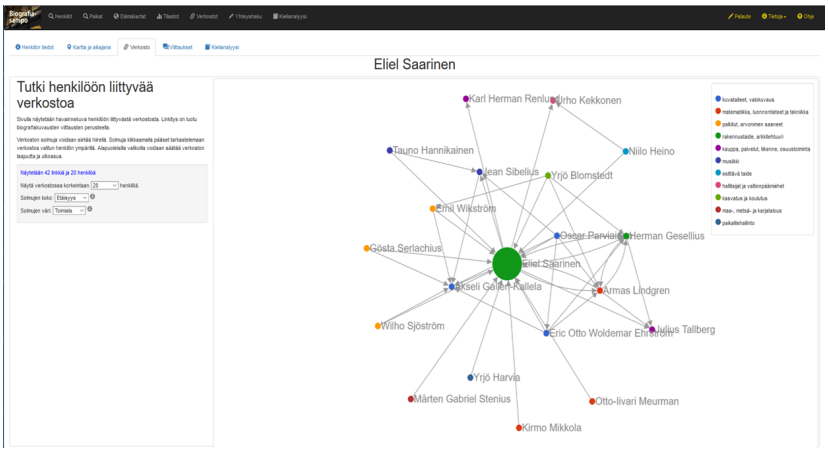


Fig. 3: Egocentric network analysis of Eliel Saarinen.

On the top of the page, there are five tabs providing data-analytic views of Saarinen. For example, Fig. 3 presents his egocentric network based on the links between the bios in the NBF, with a coloring scheme indicating persons of different types. The depth and other parameters of the network can be controlled by the widgets on the left. In Fig. 4, another tab visualizes the international events of four types of Saarinen’s life on a map and a timeline for a spatiotemporal analysis.

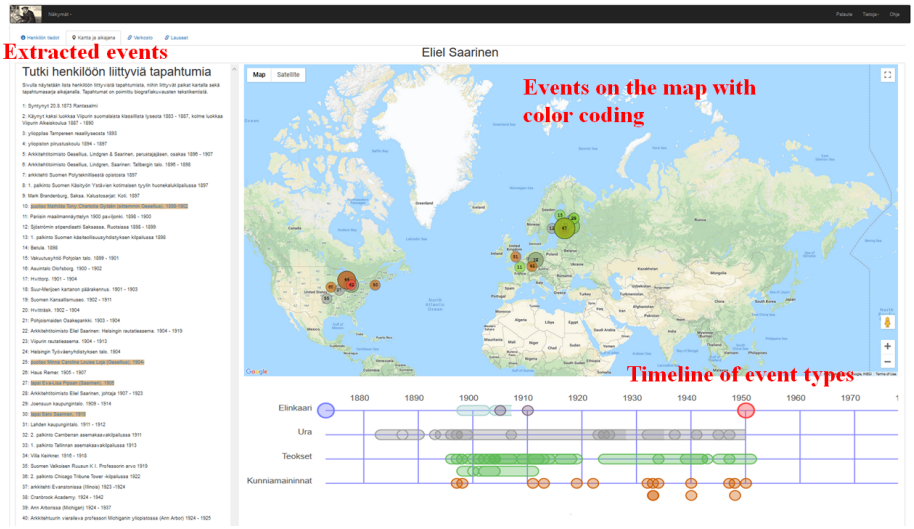


Fig. 4: Spatiotemporal visualization of the events in Eliel Saarinen's life.

2. Places BIOGRAPHYSAMPO also provides the user with a map search view in which the events extracted from the biographies are projected on the places where they occurred. After finding a place on the map, the place can be clicked. This opens a window showing the events with links to biographies. The maps in this view are not only contemporary ones but also historical maps served by the Finnish Ontology Service of Historical Places and Maps[25] [21], using a historical map service[26] based on Map Warper[27]. Many events of Finnish history took place in the eastern parts of the country that was annexed to the Soviet Union after the Second World War. Old Finnish places there may have been destroyed, placenames have been changed, and names are now written in Russian. Using semi-transparent digitized historical maps on top of contemporary maps solves the problem by giving a better historical context for the events.

3. Life Maps This perspective contains two kind of prosopographical tools: 1) *Event maps* show how different events (births, deaths, career events, artistic creation events, and accolades) that a target group of people participated in are distributed on maps. 2) *Life charts* summarize the lives of persons from a transitional perspective as blue-red arrows from the birth places (blue end) to the places of death (red end).

The prosopographical tools and visualizations in BIOGRAPHYSAMPO can be applied not only to one target group but also to two parallel groups in order to compare them. For example, Fig. 5 compares the life charts of Finnish generals and admirals in the Russian armed forces in 1809–1917 when Finland was an autonomous Grand Duchy within the Russian Empire (on the left) with the members of the Finnish clergy

[25] http://hipla.fi
[26] http://mapwarper.onki.fi/
[27] https://github.com/timwaters/mapwarper

(1800–1920) (on the right). With a few selections from the facets the user can see that, for some reason, quite a few soldiers moved the to south to die (like retirees today) while the Lutheran ministers tended to stay in Finland. The arrows are interactive. For example, by clicking on the peculiar upper arrow to the east, one can find out that this arrow was due to general Gustaf A. Silfverhjelm's (1799–1864) biography, where one can learn that he was promoted to become a chief cartographer in western Siberia.

4. Statistics The statistical application perspective includes histograms showing various numeric value distributions of the members of the group, e.g., their ages, number of spouses and children, and pie charts visualizing proportional distributions of professions, societal domains, and working organizations.

5. Networks The networks perspective is used for visualizing and studying networks among the target group. The networks are based on the reference links between the biographies, either handmade or based on automatically detected mentions. The depth of the networks can be controlled by limiting the number of links, and coloring of the nodes can be based on the gender or societal domain of the person (e.g., military, medical, business, music, etc.).

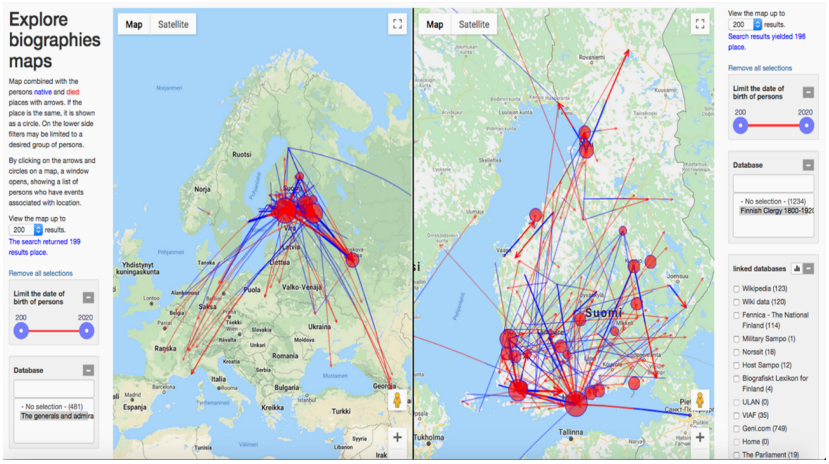


Fig. 5: Comparing the life charts of two prosopographical target groups, admirals and generals (left) and clergy (right) of the historical Grand Duchy of Finland (1809–1917).

6. Relations To utilize reasoning and knowledge discovery, an application perspective for finding "interesting/serendipitous" [3] connections in the biographical knowledge graph was created. This application idea is related to relational search [27,35]. However, in our case a new knowledge-based approach was developed to find out in what ways (groups of) people are related to places and areas. This method, described in more detail in [20], rules out non-sense relations effectively and is able to create natural language explanations for the connections. The queries are formulated and the problems are solved using faceted search. For example, the query "How are Finnish

artists related to Italy?" is solved by selecting "Italy" from the place facet and "artist" from the profession facet. The results include connections of different types (that could be filtered in another facet), e.g., that "Elin Danielson-Gambogi received in 1899 the Florence City Art Award" and "Robert Ekman created in 1844 the painting 'Landscape in Subiaco' depicting a place in Italy".

7. Language The biographies can be analyzed by using linguistic analysis, providing yet another different perspective for studying them. Both individual biographies as well as groups of them can be analyzed and compared with each other as in prosopography above. For example, it turns out that the biographies of female members of the Finnish Parliament frequently contain words "family" and "child", but these words are seldom used in the biographies of male Parliament members. The analyses are based on the linguistic knowledge graph of the texts.

Re-using the Data The different application perspectives above were implemented without modifying the data or other perspectives, but by only modifying the way the data is accessed using SPARQL.

5 Discussion

Biographical and Prosopographical Research BIOGRAPHYSAMPO offers historians and general public tools that can be used without experience in computer science. For biographical research focusing on one individual, it enriches the in-depth biographies of the NBF and offers several visualization tools. Most importantly, the portal gives scholars novel prosopographical tools for analyzing groups and networks. The tools combine quantitative approach and distant reading methods [32] with the qualitative approach, often based on close reading, typical to biographical research. BIOGRAPHYSAMPO also offers new possibilities for analyzing the language Finnish historians use in the biographies of people of different gender, age, and social groups.

BIOGRAPHYSAMPO has had 43 000 distinct users during its first five months, which indicates interest in this kind of web services. However, as of yet, the new data analytic features of the portal have not been evaluated in real-life scholarly research. We do know that the datasets and tools have certain premises and limitations that scholars have to be aware of when they use the tools. One should pay close attention to the following questions: 1) *Who created the datasets and to what end?* The core data in BiographySampo comes from biographical databases created in projects carried out by the Biographical Centre of the Finnish Literature Society in co-operation with several learned societies: the Finnish Historical Society, The Finnish Economic History Association, and the Finnish Society of Church History. This good, academically sound information has been enriched with web resources such as Wikipedia and genealogical sites like Geni.com where everyone can contribute. In BiographySampo, the source of information is always indicated – it may not be of interest to most users, but for scholars it is essential. 2) *How was the biographical collection constructed?* When it comes to biographical collections such as the NBF, the construction of the collection and the process and criteria of inclusion and exclusion of historical persons is vital information. Without understanding the process, we cannot understand who the real subject of our analysis is when we work with the datasets and tools.

BIOGRAPHYSAMPO includes two different types of biographical datasets: Firstly, there are historical groups that have been recognized by their members and outsiders as a distinct group in a given time in history, e.g., the Lutheran ministers of the Diocese of Turku in the dataset Finnish Clergy 1554–1721. The dataset includes them all and thus makes true prosopographical research possible. This is where BiographySampo is at its very best. Ministers are an especially interesting group from the point of view of networks, as the vocation often went down from grandfather to father to son, and ministers often married the daughters of other clergy families. Secondly, there are groups created by historians. For example, the National Biography of Finland, or indeed of any given country, is an artificial group. In their lifetime the biographees were not connected and certainly did not identify with each other. In network analysis, for example, the egocentric network of Blanche de Namur, the Swedish queen Blanka (1318–1363), includes Albert Edelfelt (1854–1905) who lived 500 years later, because he depicted the queen in his famous painting, not because he was in the social network of the queen.

The biographies of NBF cover one thousand years and include, e.g., all Swedish kings who ruled Finland, a witch burned at stake in the 17th century, a 18th century prostitute, the first female professor in Finland, and the software engineer Linus Torvalds, the father of the Linux operating system. What all these people from different times and different walks of life do have in common is that they have been chosen by a large and authoritative group of Finnish scholars to form a biographical representation of the history of the Finns. Some of them were eminent in their own times, some represent an important group or a phenomenon, many were pioneers in their own fields.

The statistical or linguistic analysis of these artificial groups therefore tells us not about the past itself, but about the values and preferences of Finnish historians around the turn of the millennium. As an example, we compared above the language used in the biographies of male versus female Members of Parliament (MP). The results tell us very little about the MPs and their work, but illustrate how Finnish scholars emphasize different issues when writing about the work of male and female biographees.

There is still work to be done in developing BIOGRAPHYSAMPO, its tools and data, so that they can be better understood by the users, especially those who are doing serious historical research. More background information on the datasets and the collections they are based on is needed in order to make transparent how the tools process the information. Historians are trained in source criticism and used to work with complicated documents. Digital Humanities resources should take this into account and help scholars understand and critically evaluate the tools they are offered.

Related Work Aside publishing biographical dictionaries in print and on the web, representing and analyzing biographical data has grown into a new research and application field. In 2015, the first Biographical Data in Digital World workshop BD2015 was held presenting several works on studying and analyzing biographies as data [4], and the proceedings of BD2017 contain more similar works [6]. BIOGRAPHYSAMPO is a result of research in this area and is related to several other works. In [25], analytic visualizations were created based on U.S. Legislator registry data. The idea of biographical network analysis is related to the Six Degree's of Francis Bacon system[28] [38,24] that utilizes data of the Oxford Dictionary of National Biography. However, in our case

[28] http://www.sixdegreesoffrancisbacon.com

faceted search can be used for filtering and studying target groups. The work on BIOGRAPHYSAMPO was influenced by the early Semantic NBF demonstrator [14] and its follow-up prototype [19], whose software has been applied also to a historical register of students [18] and to the U.S. Legislator data [29]. However, BIOGRAPHYSAMPO extends these systems into several new directions in terms of the DH tooling provided, such as faceted network analysis views, relational search, and text analysis views for studying the language of the biographies. Also more heterogeneous datasets are used.

Extracting RDF and OWL data from natural language texts has been studied in several works in semantic web research, cf. e.g. [8]. In [7] language technology was applied for extracting entities and relations in RDF using Dutch biographies as data in the BiographyNet. This work was part of the larger NewsReader project extracting structured data from news [31]. This line of research is similar to ours, based on the idea of extracting semantic RDF structures from unstructured biographical texts, and using the data for DH research in biography and prosopography. However, the work on BiographyNet focuses more on challenges of natural language processing and managing the provenance information of data from multiple sources, while the focus of BIOGRAPHYSAMPO is on providing the end user, both DH researchers and the general public, with intelligent search and browsing facilities, enriched reading experience, and easy to use data-analytic tooling for biography and prosopography. In addition and in contrast to the related works, BiographySampo employs the "Sampo" model [17], where the data is enriched through a shared content infrastructure by related external heterogeneous datasets, here, e.g., collection databases of museums, libraries, and archives, a critical edition, genealogical data, and various biographical data sources and semantic portals online. BIOGRAPHYSAMPO is a step in the Sampo series of semantic portals including also CultureSampo (2009), TravelSampo (2011), BookSampo (2011) (2 million users in 2018), and WarSampo (2015) (230 000 users in 2018), and NameSampo (tens of thousands of users in 2019).

Conclusions This paper presented and demonstrated the vision of a paradigm shift in publishing and using biography collections as Linked Data. The vision has been implemented as the semantic portal BIOGRAPHYSAMPO now in use on the Web. The legacy biography publishing system of SKS has 300 000 annual users on the web. We expect more users for BIOGRAPHYSAMPO since it provides the user with all biographies of SKS openly without a pay wall, the intelligent search, browsing, and DH services presented in this paper, and lots of additional enriching content interlinked from external data sources. The data of the portal was extracted and aggregated automatically by the computer. The biographical and prosopographical data-analytic tools on top of the LD service combine quantitative approach and distant reading methods [32] with the qualitative approach, traditionally based on close reading in biographical research.

Acknowledgements Thanks to Business Finland for financial support and CSC – IT Center for Science, Finland, for computational resources.

References

1. J. V. Snellman: Kootut teokset 1–24. Ministry of Education and Culture, Helsinki (2002)
2. Suomen kansallisbiografia 1–10. Suomalaisen Kirjallisuuden Seura, Helsinki (2003)

3. Aylett, R.S., Bental, D.S., Stewart, R., Forth, J., G.Wiggins: Supporting serendipitous discovery. In: Digital Futures (Third Annual Digital Economy Conference), 23–25 October, 2012, Aberdeen, UK (2012), http://www.serena.ac.uk/papers/sthash.2aHjBNNz.dpuf
4. ter Braake, S., Fokkens, A., Sluijter, R., Declerck, T., Wandl-Vogt, E. (eds.): BD2015 Biographical Data in a Digital World 2015. CEUR Workshop Proceedings, Vol. 1272 (2015)
5. Doerr, M.: The CIDOC CRM—an ontological approach to semantic interoperability of metadata. AI Magazine 24(3), 75–92 (2003), https://doi.org/10.1609/aimag.v24i3.1720
6. Fokkens, A., ter Braake, S., Sluijter, R., Arthur, P., Wandl-Vogt, E. (eds.): BD2017 Biographical Data in a Digital World 2015. CEUR Workshop Proceedings, Vol-1399 (2017), http://ceur-ws.org/Vol-2119/
7. Fokkens, A., ter Braake, S., Ockeloen, N., Vossen, P., Legêne, S., Schreiber, G., de Boer, V.: Biographynet: Extracting relations between people and events. In: Europa baut auf Biographien. pp. 193–224. New Academic Press, Wien (2017)
8. Gangemi, A., Presutti, V., Recupero, D.R., Nuzzolese, A.G., Draicchio, F., Mongiovì, M.: Semantic web machine reading with fred. Semantic Web Journal 8, 873–893 (2017)
9. Gardiner, E., Musto, R.G.: The Digital Humanities: A Primer for Students and Scholars. Cambridge University Press, New York, NY, USA (2015)
10. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool (2011), http://linkeddatabook.com/editions/1.0/
11. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating NLP using linked data. In: International semantic web conference. pp. 98–113. Springer (2013)
12. Hyvönen, E.: Preventing interoperability problems instead of solving them. Semantic Web Journal 1(1–2), 33–37 (December 2010)
13. Hyvönen, E.: Publishing and using cultural heritage linked data on the semantic web. Morgan & Claypool, Palo Alto, CA (2012)
14. Hyvönen, E., Alonen, M., Ikkala, E., Mäkelä, E.: Life stories as event-based linked data: Case semantic national biography. In: Proceedings of ISWC 2014 Posters & Demonstrations Track. CEUR Workshop Proceedings, Vol. 1272 (October 2014)
15. Hyvönen, E., Mäkelä, E., Kauppinen, T., Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Takala, J., Puputti, K., Kuittinen, H., Viljanen, K., Tuominen, J., Palonen, T., Frosterus, M., Sinkkilä, R., Paakkarinen, P., Laitio, J., Nyberg, K.: CultureSampo – Finnish culture on the Semantic Web 2.0. Thematic perspectives for the end-user. In: Museums and the Web 2009, Proceedings. Archives and Museum Informatics, Toronto (2009)
16. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers. pp. 226–230. Springer–Verlag (May 2014)
17. Hyvönen, E.: Cultural heritage linked data on the semantic web: Three case studies using the Sampo model. In: VIII Encounter of Documentation Centres of Contemporary Art: Open Linked Data and Integral Management of Information in Cultural Centres Artium, Vitoria-Gasteiz, Spain, October 19-20, 2016 (2016), https://seco.cs.aalto.fi/publications
18. Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., Sirola, L.: Reassembling and enriching the life stories in printed biographical registers: Norssi high school alumni on the semantic web. In: Language, Technology and Knowledge. pp. 113–119. Springer–Verlag (2017)
19. Hyvönen, E., Leskinen, P., Tamper, M., Tuominen, J., Keravuori, K.: Semantic National Biography of Finland. In: Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018). pp. 372–385. CEUR Workshop Proceedings, Vol-2084 (2018)
20. Hyvönen, E., Rantala, H.: Knowledge-based relation discovery in cultural heritage knowledge graphs. In: Proceedings of the 4th Digital Humanities in the Nordic Countries Conference (DHN 2019). CEUR Workshop Proceedings (2019)

21. Ikkala, E., Tuominen, J., Hyvönen, E.: Contextualizing historical places in a gazetteer by using historical maps and linked data. In: Proceedings of DH 2016. pp. 573–577 (2016)
22. Keith:, T.: Changing conceptions of National Biography. Cambridge University Press (2004)
23. Koho, M., Heino, E., Hyvönen, E.: SPARQL Faceter—Client-side Faceted Search Based on SPARQL. In: Joint Proc. of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop. CEUR Workshop Proceedings, Vol. 1615 (2016)
24. Langmead, A., Otis, J., Warren, C., Weingart, S., Zilinski, L.: Towards interoperable network ontologies for the digital humanities. Int. J. of Humanities and Arts Computing 10 (2016)
25. Larson, R.: Bringing lives to light: Biography in context. Final project report (2010), http://metadata.berkeley.edu/Biography_Final_Report.pdf, University of Berkeley
26. Le Boeuf, P., Doerr, M., Ore, C.E., Stead, S. (eds.): Definition of the CIDOC Conceptual Reference Model, Version 6.2.4. ICOM/CIDOC Documentation Standards Group (CIDOC CRM Special Interest Group) (2018), http://www.cidoc-crm.org/Version/version-6.2.4
27. Lohmann, S., Heim, P., Stegemann, T., Ziegler, J.: The RelFinder user interface: Interactive exploration of relationships between objects of interest. In: Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI 2010). pp. 421–422. ACM (2010), http://doi.acm.org/10.1145/1719970.1720052
28. Mäkelä, E., Ruotsalo, T., Hyvönen: How to deal with massively heterogeneous cultural heritage data—lessons learned in CultureSampo. Semantic Web 3(1) (2012)
29. Miyakita, G., Leskinen, P., Hyvönen, E.: Using linked data for prosopographical research of historical persons: Case U.S. Congress Legislators. In: 7th International Conference, EuroMed 2018, Nicosia, Cyprus. Springer-Verlag (2018)
30. Roberts, B.: Biographical Research. Understanding social research, Open University Press (2002), https://books.google.fi/books?id=04ScQgAACAAJ
31. Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., Bogaard, T.: Building event-centric knowledge graphs from news. Web Semantics: Science, Services and Agents on the World Wide Web 37, 132–151 (2016)
32. Schoultz, A., Matteni, A., Isele, R., Bizer, C., Becker, C.: LDIF – linked data integration framework. In: Proceedings of the 2nd International Workshop on Consuming Linked Data (COLD 2011). CEUR Workshop Proceedings, Vol. 782 (2011)
33. Shultz, K.: What is distant reading? New York Times (June, 24, 2011), https://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html accessed: 13 August 2018
34. Tamper, M., Leskinen, P., Apajalahti, K., Hyvönen, E.: Using biographical texts as linked data for prosopographical research and applications. In: Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. 7th International Conference, EuroMed 2018, Nicosia, Cyprus. Springer–Verlag (November 2018)
35. Tartari, G., Hogan, A.: WiSP: Weighted shortest paths for RDF graphs. In: Proceedings of VOILA 2018. CEUR Workshop Proceedings, Vol. 2187 (2018)
36. Tuominen, J., Hyvönen, E., Leskinen, P.: Bio CRM: A data model for representing biographical data for prosopographical research. In: Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017). vol. 2119, pp. 59–66. CEUR Workshop Proceedings (2018), http://ceur-ws.org/Vol-2119/paper10.pdf
37. Verboven, K., Carlier, M., Dumolyn, J.: A short manual to the art of prosopography. In: Prosopography approaches and applications. A handbook, pp. 35–70. Unit for Prosopographical Research (Linacre College) (2007)
38. Warren, C., Shore, D., Otis, J., Wang, L., Finegold, M., Shalizi, C.: Six degrees of Francis Bacon: A statistical method for reconstructing large historical social networks. Digital Humanities Quarterly 10 (2016), http://digitalhumanities.org/dhq/vol/10/3/000244/000244.html
39. Wu, Y., Sun, H., Yan, C.: An event timeline extraction method based on news corpus. In: 2017 IEEE 2nd International Conference on Big Data Analysis. pp. 697–702. IEEE (2017)

Publication VI

Petri Leskinen and Eero Hyvönen. Extracting Genealogical Networks of Linked Data from Biographical Texts. In *The Semantic Web: ESWC 2019 Satellite Events*, Pascal Hitzler, Sabrina Kirrane, Olaf Hartig, Victor de Boer, Maria-Esther Vidal, Maria Maleshkova, Stefan Schlobach, Karl Hammar, Nelia Lasierra, Steffen Stadtmüller, Katja Hose, Ruben Verborgh (editors), June 2019, pages 121–125, Springer, ISBN 978-3-030-32327-1, online https://doi.org/10.1007/978-3-030-32327-1_24.

Extracting Genealogical Networks of Linked Data from Biographical Texts

Petri Leskinen[1][0000−0003−2327−6942] and Eero Hyvönen[1,2][0000−0003−1695−5840]

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

Abstract. This paper presents the idea and our work of extracting and reassembling a genealogical network automatically from a collection of biographies. The network can be used as a tool for network analysis of historical persons. The data has been published as Linked Data and as an interactive online service as part of the in-use system *BiographySampo – Finnish Biographies on the Semantic Web*.

1 Introduction

Extracting and inferring social or genealogical networks from historical documents can provide new information for biographical and prosopographical [1] research. However, genealogical data is often available only in textual form providing challenges for knowledge extraction: How to identify persons and their gender by different name forms? How to disambiguate namesakes in different times? How to extract the genealogical relations between the mentions? This paper presents a case study for extracting the explicit genealogical network implicit in the national collection of 13144 Finnish biographies[3]. The methodological idea is to combine regular expression identification, imprecise proper name matching, gender information, and data about expected lifespans for more accurate results. The system was evaluated with promising results, and a tool was constructed, based on Linked Data, for examining the underlying network of ∼81 000 extracted basic relations 'mother', 'father', 'wife', 'husband', 'son', and 'daughter'. On top of the Linked Data service, a new application was created for studying the networks interactively as a new part of the in-use BiographySampo[4] system [2].

Related work Extracting biographical networks are discussed in *Six Degrees of Francis Bacon* [3]. Articles [4,5] discuss extracting genealogical networks from multi-source vital records. For the large public there are many crowd-sourcing-based commercial genealogy websites, such as *ancestry.com*, *myheritage.com*, and *geni.com*. This paper extends our earlier papers about BiographySampo [2] and network analysis based on biographical link references into extraction of genealogical networks, and presents a new visual application view for studying such networks interactively.

[3] https://kansallisbiografia.fi/, accessed 20 March 2019
[4] http://biografiasampo.fi

2 Extracting Genealogical Networks from Texts

Dataset BiographySampo is a semantic portal based on a knowledge base that has been created using natural language processing methods, linked data, and semantic web technologies. It contains 13 144 biographies of notable Finns that can be browsed through a faceted search application and using tools for Digital Humanities research. [6] In addition to the genealogical network discussed in this paper, the data been a source for reference network extraction [7,8].

Pattern-based Knowledge Extraction Many biographies in the dataset include semi-formal textual descriptions of family relations of the protagonist. As an example, the description of baroness *Elisabeth Järnefelt*[5] is given below:

> Jelizaveta Konstantinovna Clodt von Jürgensburg from year 1857 known as Järnefelt, Elisabeth S 11.1.1839 Pietari, K 3.2.1929 Helsinki.
> *V* Baron, major general Konstantin Karlovitsh Clodt von Jürgensburg and Catharine Vign. *P* 1857 senator, governor, lieutenant general August Alexander Järnefelt *S* 1833, *K* 1896, *PV* bailiff Gustaf Adolf Järnefelt and Aurora Fredrika Molander.
> *Children:* Caspar (Kasper) Woldemar *S* 1859, *K* 1941, critic, translator, Russian language teacher, painter, *P* Emma Ahonen; Edvard Armas *S* 1869, *K* 1958, conductor, composer, professor, *P1* songstress Maikki Pakarinen, *P2* songstress Olivia (Liva) Edström; Aina (Aino) *S* 1871, *K* 1969, *P* composer Jean Sibelius;

The semi-formal expressions here have uniformity in structure that can be used effectively for pattern-based information extraction: First, the given and family names are mentioned and after that the years of birth S and death K. The description provides information about the parents (marked with V), spouses (P), parents-in-law (PV), children, and children-in-law of the protagonist.

One major problem in knowledge extraction here is recognizing the same person, here *Elisabeth Järnefelt*, referenced with different names: *Jelizaveta Konstantinovna Clodt von Jürgensburg*, *Elisabeth Clodt von Jürgensburg* or most commonly with *Elisabeth Järnefelt*. On the other hand, same names are used in families over and over again. For example, there is a case of four people with name *Christian Trapp*, a grandfather, a father, a son[6], and a grandson. They cannot be distinguished without additional information about their known lifespans.

Data Processing In our knowledge extraction pipeline, the genealogical textual description of the protagonist is first divided into the parts describing his/her parents, spouses (wife/husband distinction is not known at this point), and children. The division is based on using regular expressions matching the punctuation and the tokens V, P, PV. The years of birth, death, or marriage are easily separated from the text sequence. To separate occupational descriptions from the proper names, we used the ARPA service[7] together with vocabularies of Finnish female, male, and family names[8].

The extracted names were used to reason the gender of the person, which was used to refine relations, e.g., to specify a *parent* as a *mother* or a *father*. For

[5] `http://biografiasampo.fi/henkilo/p3148`
[6] `http://biografiasampo.fi/henkilo/p10013`
[7] `http://seco.cs.aalto.fi/projects/dcert/`, accessed: 9 March 2019
[8] `https://www.avoindata.fi/data/en_GB/dataset/none`, accessed: 20 March 2019

the network, the spouses were linked with the children by the known years of marriage and child birth.

To gain detailed vital information for disambiguation, we reasoned lifetime estimates, e.g., the missing years of birth of the parents based on the known birth year of their child. The estimates were constructed by first collecting the years of births of a parent and a child from the known cases in data. The distributions of parent ages at child birth are depicted in Fig. 1. To reason the ages of spouses, a similar study was performed with the result that 99% of differences between the births of a husband and wife is in the range of -18–+35 years. The more relatives with known vital records a person has, the more precise the estimates are.

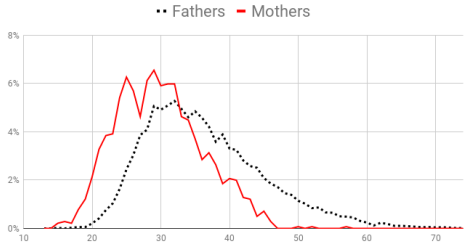


Fig. 1: Distribution of parent ages at a child birth

3 Evaluation

For evaluation we chose 50 random biographies, and manually compared the texts with the extracted results. The test set had mentions of 170 people. We compared the data fields of person names, years of birth and death, gender, occupation, and relation type. According to our evaluation 94.5% of the extracted people records were mentioned only in a single biography. The accuracy for these people was 97.3%, and 80.4% for people mentioned in multiple biographies. The system for inferring the gender could recognize 97.7% of the names leaving out very rare or foreign names. In our test set all inferred genders were correct.

For an example of the extracted network, the genealogical network of Elisabeth Järnefelt[9] is (partly) depicted in Fig. 2. She turns out to be a part of the largest connected subnetwork in our data. This subnetwork has 2694 family relation links and connects 1835 people mentioned in 250 biographies.

To further enrich the web portal, the network of immediate family members was used to reason other relatives of each protagonist[10]: the siblings, cousins, uncles, aunts, grandparents, grandchildren, and relatives-in-law.

[9] `http://biografiasampo.fi/henkilo/p3148/sukulaiset`
[10] `http://biografiasampo.fi/henkilo/p3148`

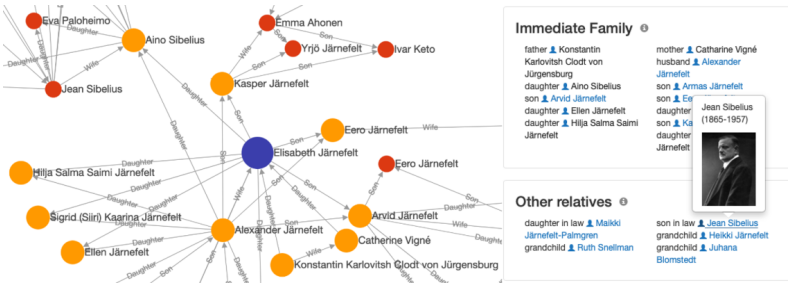


Fig. 2: Left: Genealogical network around Elisabeth Järnefelt as seen in a BiographySampo view. Right: relations shown on the web portal, including the composer Jean Sibelius, husband of Elisabeth's daughter Aino.

Acknowledgements Thanks to Business Finland for financial support and CSC – IT Center for Science, Finland, for computational resources.

References

1. Verboven, K., Carlier, M., Dumolyn, J.: A short manual to the art of prosopography. In: Prosopography approaches and applications. A handbook. Unit for Prosopographical Research (Linacre College) (2007) 35–70
2. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., Keravuori, K.: BiographySampo - publishing and enriching biographies on the semantic web for digital humanities research. In: Proceedings of ESWC 2019, Springer-Verlag (2019) Accepted.
3. Warren, C.N., Shore, D., Otis, J., Wang, L., Finegold, M., Shalizi, C.: Six degrees of Francis Bacon: A statistical method for reconstructing large historical social networks. DHQ: Digital Humanities Quarterly **10** (2016)
4. Efremova, J., Ranjbar-Sahraei, B., Rahmani, H., Oliehoek, F.A., Calders, T., Tuyls, K., Weiss, G.: Multi-source entity resolution for genealogical data. In: Population reconstruction. Springer-Verlag (2015) 129–154
5. Malmi, E., Rasa, M., Gionis, A.: AncestryAI: A tool for exploring computationally inferred family trees. In: Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee (2017) 257–261
6. Hyvönen, E., Leskinen, P., Tamper, M., Tuominen, J., Keravuori, K.: Semantic National Biography of Finland. In: Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018), CEUR Workshop Proceedings, Vol-2084 (2018) 372–385 `http://www.ceur-ws.org/Vol-2084/short12.pdf`.
7. Tamper, M., Leskinen, P., Apajalahti, K., Hyvönen, E.: Using biographical texts as linked data for prosopographical research and applications. In: Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. 7th International Conference, EuroMed 2018, Nicosia, Cyprus, Springer-Verlag (2018)
8. Tamper, M., Hyvönen, E., Leskinen, P.: Visualizing and analyzing networks of named entities in biographical dictionaries for digital humanities research. In: Proceeedings of CICLing 2019, Springer-Verlag (2019) Accepted.

Publication 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, December 2022, ISSN 1570-0844 (P), DOI 10.3233/SQ-222887, online https://doi.org/10.3233/SW-222887 .

Analyzing Biography Collections Historiographically as Linked Data: Case National Biography of Finland

Minna Tamper [a] , Petri Leskinen [a] , Eero Hyvönen [a,b] , Risto Valjus [c] , and Kirsi Keravuori [c]
[a] *Semantic Computing Research Group (SeCo), Aalto University, Department of Computer Science, Finland*
E-mail: firstname.lastname@aalto.fi
[b] *HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland*
E-mail: firstname.lastname@helsinki.fi
[c] *The Finnish Literature Society, Finland*
E-mail: firstname.lastname@finlit.fi

Abstract. Biographical collections are available on the Web for close reading. However, the underlying texts can also be used for data analysis and distant reading, if the documents are available as data. Such data is usable for creating intelligent user interfaces to biographical data, including Digital Humanities tooling for visualizations, data analysis, and knowledge discovery in biographical and prosopographical research. In this paper, we re-use biographical collection data from a historiographical perspective for analyzing the underlying collection. For example: What kind of people have been included in the collection? Does the language used for describing female biographees differ from that for men? As a case study, the Finnish National Biography, available as part of the Linked Open Data service and semantic portal *BiographySampo – Finnish Biographies on the Semantic Web* is used. The analyses show interesting results related to, e.g., how specific prosopographical groups, such as women or professional groups are represented and portrayed. Various novel statistics and network analyses of the biographees are presented. Our analyses give new insights to the editors of the National Biography as well as to researchers in biography, prosopography, and historiography. The presented approach can be applied also to similar biography collections in other countries.

Keywords: Linked Data, Data Analysis, Network Analysis, Cultural Heritage, Digital Humanities

1. Introduction

Biographical dictionaries are scholarly resources used by the public and by the academic community alike. Most national biographical dictionaries follow the traditional form of combining a lengthy non-structured text, often written with authorial individuality and personal insight, with a structured synopsis of basic biographical facts, such as family relations, education, works, career events, and so on. Biographies are an invaluable information source for researchers across various disciplines with an interest in the past. [1] A well-known example of a biographical dictionary is the Oxford Dictionary of National Biography (ODNB)[1] with more than 60 000 lives. It was published in print and online in 2004, and since then many dictionaries have opened their editions on the Web. These include USA's American National Biography[2], Austrian Prosopographical Information System[3], Germany's Neue Deutsche Biographie[4], Biography Portal of the Netherlands[5], The Dictionary of Swedish National Biography[6], and the National Biog-

[1] http://global.oup.com/oxforddnb/info/
[2] http://www.anb.org/aboutanb.html
[3] https://apis.acdh.oeaw.ac.at/
[4] http://www.ndb.badw-muenchen.de/ndb_aufgaben_e.htm
[5] http://www.biografischportaal.nl/en
[6] https://sok.riksarkivet.se/Sbl/Start.aspx?lang=en

raphy of Finland[7] (NBF). There are also many "who is who" services online, and Wikipedia contains lots of short biographies.

In this paper, we use the BiographySampo portal and its data, based on the National Biography of Finland, to study and analyze biographees, their lives, and the source material with two goals in mind. Firstly, our goal is to argue and show that using biographies as Linked Data opens up unprecedented new possibilities for the study by distant reading [3, 4]. Secondly, the analyses present novel insights into the nature and contents of the NBF. Here, our focus is on the historiographical analysis of biographies. We anticipate that comparative results can be expected, if the methodology and tools introduced are applied to similar national biographical dictionaries. Our approach can also be applied to other domains of Cultural Heritage data, such as museum collections, library catalogs, manuscripts in archives, archaeological finds, etc., as demonstrated by the Sampo series of semantic portals[8] [5].

1.1. National Biography of Finland

In Finland, the National Biography collection and several other collections of biographical and prosopographical data have been compiled and are maintained by the Finnish Literature Society (SKS)[9] established in 1831. The work has been carried out by the Biographical Centre of the SKS, now part of the society's scholarly publishing house, in collaboration with several Finnish learned societies and researchers in different fields.

The kernel of the collection is the National Biography of Finland (*Suomen kansallisbiografia* in Finnish), based on the biographies written in collaboration with the Finnish Historical Society in 1993–2001. The NBF was created for an educated reader, who is not an expert in history. Historical terms and concepts are explained, and the biographees are presented within the frame of national history. The articles have been written with a critical attitude and in accordance with sound historiographical methods. The facts and the emphasis of the articles must derive from recent research and be well argued. The NBF strives to be enjoyable and interesting reading as well as to bring new insights into the impact of individuals in history. In addition to the general reader, the NBF is also a useful handbook for researchers from all fields who are seeking reliable biographical information. The articles have been peer reviewed and contain reference to archival sources and literature.

The NBF contains 6500 lives and goes back a thousand years in history. The National Biography of Finland was one of the largest projects ever carried out in the field of history in Finland: it involved twenty historians serving in the three editorial boards (Swedish era, Russian era, and Independence era) and over 900 other scholars who wrote the biographies. The writing of the articles began in 1993 and the first articles were published online in 1997 when Finland celebrated her 80 years of independence. The majority of the biographies were written before the year 2000. Some 6 000 articles were published in print in 2003–2007 (Suomen kansallisbiografia 1–10 [2]) by the Finnish Literature Society.

Early on in the project, half of the 6 000 lives to be commissioned were allocated to the period of independence from 1917 onward. The Swedish era from the earliest decades to 1809 and the Russian era from 1809 to 1917 were each given a 25 percent of the entries.

Contrary to most national biographical dictionaries, the NBF includes people who are still alive, although most of them are already past the peak of their career and activity. The reason was the emphasis on the period of independence in the work of the editorial board. Had only deceased Finns been included, the big picture of the independence era created by the lives would have been incomplete and distorted.

In addition to the NBF, the Finnish Literature Society has also published other biographical collections, 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. The biographies have been made available also as a web service[10]. In 2018, the collections were re-published as the semantic portal *BiographySampo—Finnish biographies on the Semantic Web* [6] and it has had approximately some 40 000, end-users on the Web.

[7] http://kansallisbiografia.fi [2]
[8] https://seco.cs.aalto.fi/applications/sampo/
[9] https://finlit.fi/
[10] https://kansallisbiografia.fi/english

1.2. A Paradigm Shift in Publishing Biography Collections

BiographySampo[11] [6] is a semantic portal that is based on a knowledge graph that has been extracted automatically from textual biographies to its additional metadata. The portal has been built to help historians and scholars in biographical [7] and prosopographical research [8, 9][12]. A major novelty of BiographySampo is to provide the user with data-analytic and visualization tools for solving research problems in Digital Humanities (DH), based on Linked Data [10, 11]. The idea of publishing biographies as structured Linked Data for machines with ready-to-use tooling for humans to use in Digital Humanities research can be seen as a paradigm shift in the field of biographical publishing [6, 12]. Traditionally, biographies have been published as printed texts, in our case as a series of ten volumes [2] of nearly 10 000 pages. Then, the Web emerged as a publication channel for biographies for human consumption. In the case of the NBF, this happened already in 1997. BiographySampo demonstrates the next step ahead where the biographies are published not only as texts for close reading but also as machine "understandable" Linked Data for distant reading. This facilitates data analysis and tooling to be used for DH research, and even application of Artificial Intelligence to knowledge discovery, where the machine can help the user in finding research problems, in solving them, and in explaining the results [12].

BiographySampo is based on the Sampo model [5] that formulates the idea of aggregating and publishing distributed, heterogeneous local data sources in a global linked data service. In this way, the data of all data providers can be enriched with each other's content, by reasoning based on Semantic Web standards, and the global data can be used easily across original local data silo boundaries. This arguably creates a sustainable "business model" where every data provider wins through collaboration, and of course the end users in particular. Data alignment and linking in this approach is based on a shared global data model and a set of shared domain ontologies (places, people, etc.) that are used for describing the contents of the different data sources for semantic interoperability.

The data is searched, explored, and analyzed in a kind of standardized way with the following way. Firstly, the landing page of the portal provides the user with multiple "perspectives" for searching and exploring the underlying data. In our case, biographical data can be accessed from seven search perspectives [6]: Persons, Places, Lives on maps, Statistics, Networks, Relations, and Linguistics. Secondly, each perspective provides the end-user with a semantic faceted search engine, where the results can be filtered and found flexibly by making selections using a set of orthogonal facets (e.g., place, time, person, etc.). Thirdly, after filtering down a target set of entities of interest , the set can be analyzed and visualized using a variety of ready-to-use data-analytic tools. For example, various map- and network-based visualizations and statistics are available. Furthermore, the SPARQL endpoint of the underlying Linked Open Data service can be used for querying, analyzing, and visualizing the data in flexible ways using tools, such as Yasgui [13] for SPARQL, or Jupyter[13] and Google Colab[14] by Python scripting. In this paper, analyses by both the ready-to-use tools of the portal and by using Google Colab on the underlying SPARQL endpoint will be presented. The portal interface was developed by using the SPARQL Faceter tool [14] that has later on been developed into the full stack Sampo-UI framework [15].

1.3. Related Work

Biographical collections can be used to study the underlying historical world. However, the texts, the language used, and the biographical collection as a whole can also be studied from a different, historiographical perspective as an artifact reflecting its own time, the editorial values and biases in selecting the biographees, the authors' perspectives, and also from a linguistic points of view. Such analyses have been already made for some national dictionaries of biography, e.g., for the ODNB [16] and the Irish Ainm [17].

Christopher N. Warren claims [16] that national dictionaries of biography, such as the ODNB, speak with a double voice: they give us information about things as they happened, but are at the same time a testimony

[11] Online at www.biografiasampo.fi; see project homepage https://seco.cs.aalto.fi/projects/biografiasampo/en/ for further info and publications.

[12] Prosopography is a method that is used to study groups of people through their biographical data. The goal of prosopography is to find connections, trends, and patterns from these groups.

[13] https://jupyter.org/

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

about how a key piece of historiographical infrastructure was made. He sees the ODNB as data and, at the same time, as a historical artifact. There are also related studies using, e.g., Wikipedia articles as the data source [18, 19]. This paper presents, in the same vein, a study of the National Biography of Finland. The methods and tools created in our work for the analysis are generic and can be re-used for similar tasks based on Linked Data standards. The data and SPARQL endpoint used are available at the Linked Data Finland platform[15] [20]. The work presented is novel in its way of using Linked Data for historiographical analysis of textual biographies. It is also arguably the first historiographical analysis of the NBF collection. The data is open for further analyses for anyone on the Web.

Aside publishing biographical dictionaries in print and on the Web, representing and analyzing biographical data has grown into a new research and application field. In 2015, the first Biographical Data in Digital World workshop BD2015 was held presenting several works on studying and analyzing biographies as data [21], and the proceedings of BD2017 contain more similar works [22]. In [23], analytic visualizations were created based on U.S. Legislator registry data. The idea of biographical network analysis is related to the Six Degrees of Francis Bacon system[16] [24, 25] that utilizes data of the Oxford Dictionary of National Biography. However, a novelty of our approach is to use faceted search for filtering out target groups for studying. The work was influenced by the early Semantic NBF demonstrator [26] and its follow-up prototype [27], whose software has been applied also to a historical register of students [28] and to the U.S. Legislator data [29]. However, BiographySampo extends these systems into several new directions in terms of the DH tooling provided, such as faceted network analysis views, relational search, and text analysis views for studying the language of the biographies. Also, more heterogeneous datasets are used.

Extracting Linked Data from texts has been studied in several works, cf. e.g. [30, 31]. In [32] language technology was applied for extracting entities and relations in RDF using Dutch biographies in the BiographyNet[17]. This work was part of the larger NewsReader project[18] extracting data from news [33]. This line of research is similar to ours, based on the idea of extracting RDF data from unstructured biographical texts. However, BiographyNet focuses more on the challenges of natural language processing and managing the provenance information of data from multiple sources, while our focus is on providing the end user with intelligent search and browsing facilities, enriched reading experience, and easy to use data-analytic tooling for biography and prosopography. The Austrian Prosopographical Information System (APIS) [34–36] is a virtual research environment that transforms text collections to machine readable formats and enables the use of natural language processing based methods to enrich the documents by extracting and linking information in them. The system has been used to transform and to study the collection of Austrian Biographical Dictionary 1815—1950 (ÖBL). Similarly to BiographySampo, the APIS can be used to analyze and visualize datasets using for example network analysis methods.

This paper is structured as follows. First, an overview of the NBF data and its transformation into Linked Open Data is described. After this, various data analyses are presented and discussed using the tools of the portal as well as Google Colab scripting. Finally, issues related to data quality and interpretation of the analyses are discussed, and directions for further research are outlined.

2. Transforming Biographies into Linked Open Data

This section explains contents of the NBF data to be used in our analyses, and how the source data was transformed into Linked Data and published in a SPARQL endpoint on the Semantic Web.

2.1. Source Data

BiographySampo contains some 13 100 biographies including the core NBF and four supplement datasets: Finnish Clergy 1554–1721, Finnish Clergy 1800–1920, Finnish Generals and Admirals 1809–1917, and Business Leaders. The NBF alone contains 6478 entries, 5268 men, 929 women, 11 couples, and 268 families. [37] In the NBF dataset, there were also two individual biographees whose gender is missing in the data. The earliest biographee is a saint approximately from the year 200, whereas there are also many biographies about living persons in the collection, such as

[15] http://www.ldf.fi/dataset/nbf
[16] http://www.sixdegreesoffrancisbacon.com
[17] http://www.biographynet.nl/
[18] http://www.newsreader-project.eu/

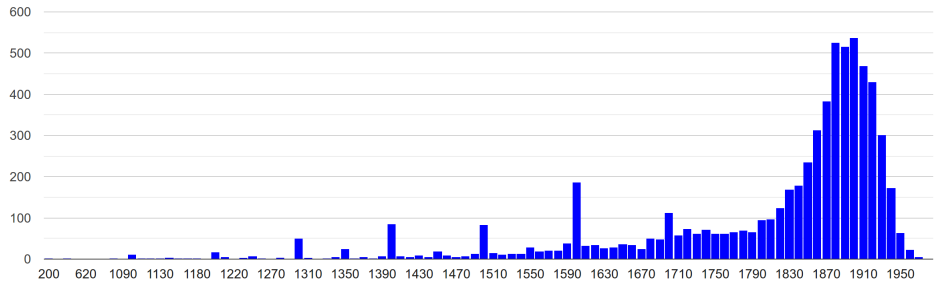


FIG. 1.: Amount of biographies by biographee's birth decade; screenshot from the BiographySampo portal

Jenni Haukio, the current First Lady of Finland. The distribution of the biographical texts by decade can be seen in Fig. 1. In this paper, only men and women in the core NBF dataset are considered; the couples and the families are left out as well as the other four supplement datasets mentioned above.

A biography text in the NBF is represented in two major parts: First, there is a narrative text on the life of the biographee, including a lead section. This text is written in ordinary natural Finnish. The text is used in the online version of the NBF and includes hand coded HTML links to related biographies in the collection; this is the only semantic markup in the text. After the free text section, a summary of the person's life is presented including basic data about the biographee (name, birth, death etc.) and information about family relations, life events, and career achievements [38]. In the NBF, the summary is unstructured text, too, but written in a semi-formal language using different section headings and notations for separating, e.g., information about family relations from career achievements. The sentences in the semi-formal part are shortened, use specific short hand notations, and do not, e.g., have predicates.

In addition to the biographical text, the NBF data includes structured metadata about the biographies and the biographees available as a spreadsheet in CSV format. The metadata contains the basic biographical information of the biographee, i.e., person names with possible variations like maiden or altered names, places and times of birth and death, vocational/occupational group of the person (Politics, Economics, Science, etc.), and a link to the photo of the person. The metadata is used as the basis for searching biographies in the online version of the NBF. In addition to biographical metadata, the dataset included information about the authors of the biographies, their gender and birth year.

In addition to the biographies, BiographySampo also makes use of several external data sources for enriching the data. For example, the biographees are linked with *same as* links to 16 additional data sources on the Web. One application perspective in BiographySampo, Relational Search for knowledge discovery [39], makes use of additional datasets extracted from collections of museums, libraries, and archives. This supplementary data is not considered or used in the analyses of this paper.

2.2. *Transformation into Linked Data*

In BiographySampo, the metadata CSV as well as the textual biographies were analyzed and transformed automatically into linked data, and links to external data sources were established. The modeling choices, transformation, and enriching of the data have been described in various articles throughout the project [37, 39–42]. The result was published as a SPARQL endpoint that was used as the basis for the semantic portal and the analyses presented in this paper. The data in the service can be divided into the following conceptual categories:

Basic information about the biographees. This data is based on the metadata CSV. A custom NBF namespace is used in addition with Dublin Core Metadata Initiative (DCMI) Metadata Terms[19] and Schema.org[20]. During the data transformation, the literal property values of persons, such as variations of family and given names, lifetime dates, and URLs for person images where transformed into data resources according to the data schema while some data values, such as vocations, vocational groups, and places of birth and death, were aligned with the domain ontologies of BiographySampo. This data is reliable as it is hand coded by the editors and authors of the NBF,

[19] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[20] https://schema.org/

and the terminology used, such as vocational groups, is controlled and unambiguous.

Metadata about biography documents. The author and publishing date data was extracted from the hand coded CSV metadata. Here, the NBF namespace is supplemented with the Dublin Core (DC) Metadata Element Set[21], DCMI Metadata Terms, and Schema.org. The free text and semi-formal summary paragraphs were categorized based on content to be able to target different categories for different data analytical applications and knowledge extraction. The content types included free text paragraphs such as the lead paragraph and the narrative text whereas the semi-formal was typed to summary of person's life, family relations, life events, and career achievements. This was done to distinguish the content type for automatic annotation processes. The lead paragraph was found from 6500 biographies, narrative text from 6500 and family relations from 6220, and career events or achievements from 6430 biographies. The accuracy of the classification of the text paragraphs was 98.5%. It was estimated for 200 randomly picked paragraphs and the most common error was mixing lead paragraph and narrative text paragraph in biographies that had unusual document structure. In addition, the subject matter of biography texts, based on the free text parts, was analyzed using automatic annotation and represented using keywords taken from the Finnish General Ontology YSO[22].

Reference network to other biographees within the NBF. The data about the biographee resources was enriched with internal links to other biographees. The links were extracted in two different ways: 1) Linkage based on the hand coded directed HTML reference links between the biographies. 2) Linkage based on mentions of persons in the free text parts of the biographies. The HTML links were extracted while transforming the text to RDF [40] with 99.4% accuracy that was estimated for randomly selected 36 documents containing 176 links. The mentioned people were extracted by the machine using automatic Named Entity Linking [41, 43]. The accuracy of named entity linking succeeded with 74.0% accuracy. The networks based on link types 1 and 2 can be used independently from each other in analyses; the choice can be made, e.g., in the portal user interface. The modeling choices are described in more detail in [40, 41].

[21] https://www.dublincore.org/specifications/dublin-core/dces/
[22] https://finto.fi/yso/en/

Linkage network to persons in external data sources. Data about the person resources was enriched with "same as" links to 16 external biographical data sources, such as Wikidata[23], Getty Union List of Artist Names (ULAN)[24], The Virtual International Authority File (VIAF)[25], Finnish databases providing biographical information, and other Sampo portals on the Semantic Web. In most cases, this linking could be made accurately using names and dates of birth and death. In addition, most of the biographees have an entry in Wikidata, especially those who lived after the 18th century. However, for people of medieval times the available information about his/her years of living might be inadequate. Different databases often use different name variations of the same person. For example, the names of notable medieval Swedish people are translated to Finnish in the NBF.

Personal life events. The life of each biographee was described semantically in terms of spatio-temporal events in which they participated in. The event data was extracted from the semi-formal summaries of the biographies using regular expressions. However, the events of birth and death are based on the CSV metadata. The life event data has been modelled using an actor-event schema based on the CIDOC CRM standard[26]. Here life events fall in different subclasses and are characterized by properties that tell the place, time, and participants of the event. According to our evaluation 97.5% of the expressions of time were correctly extracted and interpreted from the texts. The main disambiguation and linking challenge here were the historical place names used in descriptions, but this could also be performed fairly reliably with a precision of 98.4% and a recall of 85.7%.

Genealogical network. A separate genealogical network was created automatically based on the mentions of different family relations, *mother*, *father*, *child*, or *spouse* in the semi-formal part of the biographies. This data was enriched by reasoning the gender of mentioned persons if needed [44] and by inferring additional relations, such as *grandfather* or *cousin*. The genealogical network includes lots of historical persons that do not have a biography in the NBF. Generally, according to our evaluation 93.9% of the mentioned person names were correctly interpreted in our conversion process.

[23] https://www.wikidata.org/wiki/Wikidata:Main_Page
[24] https://www.getty.edu/research/tools/vocabularies/ulan/
[25] http://viaf.org/
[26] http://www.cidoc-crm.org/

Family relations are modelled using the Bio CRM model [45], an extension of the CIDOC CRM standard. The method and process of extracting the family relations is described and the results are evaluated in [42].

Linguistic descriptions of biography texts. A linguistic knowledge extraction pipeline was created for analyzing the free text parts of the biographies. It identifies text structures, such as paragraphs, sentences, and words, including morphological analysis data (e.g., part-of-speech tags (POS), lemmas, and dependency grammar information). The results were described using mainly the NLP Interchange Format (NIF) [46–48] and the CoNLL namespace by using the CoNLL-RDF [49] tool. The model was extended with the DC Metadata Element Set, DCMI Metadata Terms, and the NBF namespace for describing, for example, relations between text structures (e.g., documents and its paragraphs, sentences, and words) to facilitate querying the linguistic data in detail. The linguistic knowledge graph was also enriched with additional precalculated relations that are used for making SPARQL queries simpler and more efficient in the BiographySampo portal. According to our evaluation the linguistic graph for the NBF extraction succeeded with 100% for paragraphs, 99.5% for sentences, 99.0% for words, and 95.6% for POS tags. The results were calculated for 200 randomly selected entities in each category. Sometimes initials (e.g., J. A. von Essen) caused issues with sentence splitting and for POS tagging (the tags for initials varied between SYM and PROPN), while sometimes timespans (e.g., 2008-2009 was occasionally split to two word tokens as hyphen was included in either of the numbers) cause issues for word classification.

The quality of the data in these categories in terms of uncertainty, incompleteness, and errors is different depending on the data source and the knowledge extraction process used. This matter will be discussed later in chapter 3 when presenting and interpreting the analyses made using these data.

The final outcome of the knowledge extraction process is illustrated in Fig. 2. The linked data is divided into mutually related biographical and linguistic knowledge graphs. The size on the knowledge graphs is documented in terms of the number of instances in different classes, except for the values of LOD cloud links and Morphological data, which are amounts of triples. For example, the biographees were involved in all together 117 000 events during their lives, and the free text parts contain nearly 7 million words.

2.3. Linked Open Data Service

Finally, the transformed knowledge graphs were published openly (under the CC BY 4.0 license[27], excluding data about the biographical texts and living people) on the Linked Data Finland platform LDF.fi[28] [20]. LDF.fi provides the user with a standard SPARQL endpoint for querying the data[29], on top of which the online BiographySampo portal was implemented. In addition, the data service supports best practices on W3C for publishing Linked Data [10]. A URI identifier resolving mechanism is provided. This means, for example, that if a URI is typed in a browser, a HTML protocol is returned that shows the corresponding data as a human readable HTML page that can be inspected and browsed further by linked data browsing. In the same vein, the data in RDF form can be accessed by applications by using the HTML protocol. It is also possible to download the data in textual form for off-line processing. The LDF.fi platform also includes additional tools that aim at helping the user to re-use the data. For example, schemas are documented automatically for the human user by a schema documentation generator, the LODE Documentation Environment[30] service. The data model for the NBF is documented for people and biography metadata in [6], linguistic knowledge graph in [40], and for enrichment with named entities in [41].

3. Analyzing and Visualizing the National Biography of Finland

In this chapter, we present analyses based on the NBF data service. In BiographySampo there are ready-to-use tools [40, 42, 50] for general statistics and more conceptual categories such as linguistic analysis, network analysis, and map visualizations. This chapter starts with general statistics. After this more detailed analyses based on the conceptual categories of data are presented and interpreted. Some analyses can be tested online in BiographySampo as part of the tool set available there. For others, the SPARQL endpoint has been used with Google Colab, and a variety of Python data analysis and visualization tools such as Matplotlib[31].

[27] https://creativecommons.org/licenses/by/4.0/

[28] https://ldf.fi

[29] See the dataset home page at https://www.ldf.fi/dataset/nbf for more details.

[30] https://essepuntato.it/lode/

[31] https://matplotlib.org/

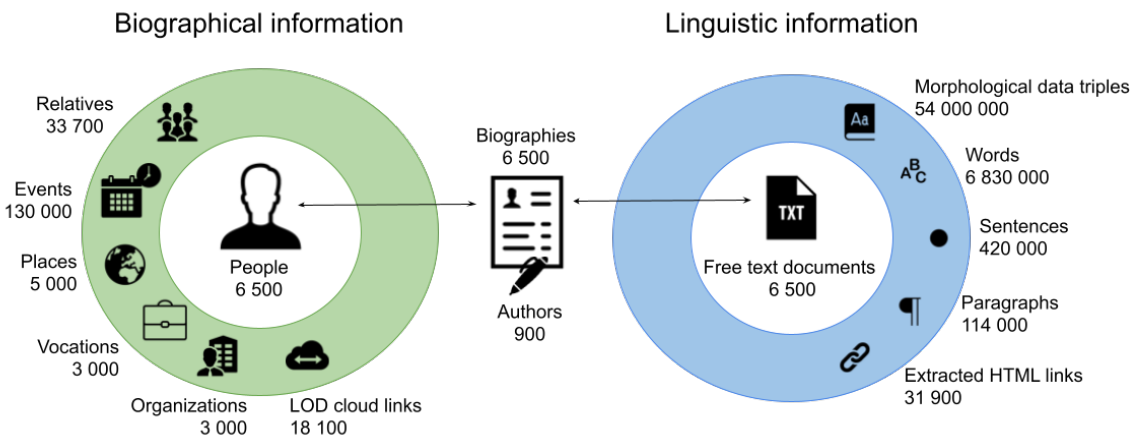


FIG. 2.: Amounts of extracted biographical and linguistic data.

3.1. General Collection Statistics

The general statistics of the NBF can be created and visualized in BiographySampo with versatile options. The statistics tell about the demographic nature of the people included in the dataset. The statistical tools are available online through a "Statistics" application perspective[32], with separate tabs for histograms, pie chars, and a Sankey chart for analyzing the family relations of the biographees. In all tabs it is possible to focus the statistical analyses prosopographically to subsets of biographees, such as women or people born on a certain time period in Helsinki, by using a faceted search/filtering engine. Filtering the data is also possible using non-demographic metadata, such as authorship of the biographies and the inclusion of the biographee in other data sources, such as Wikipedia/Wikidata or ULAN. In addition, there are separate tabs available for making comparisons between subsets of the biographees, like between two vocational groups.

In Fig. 1, the number of biographies have been plotted by decade. The plot is taken from the BiographySampo portal's statistical analysis page. In the plot, the decade has been selected based on the birth year of the biographee. The distribution shows a peak of biographies that have been written about people born between the end of 19th century and the beginning of the 20th century and they have been active when the Finnish identity as a sovereign nation was established. There are also a few peaks earlier in history that are in general less well-known in Finnish history. In some cases, the data is not accurate enough and the birth year of a biographee is not known. In these cases it has been set to the beginning of a century, which explains the earlier peeks in the beginning of each century.

Similarly to [16] we have plotted the distribution of people alive on a timeline. Based on biographee's birth and death data. Figure 3 depicts the number of biographees alive in different times but due to lack of total population information in Finland before 1900s we do not have comparison between biographees and general population but we wanted to look at women in contrast to all biographees. The blue curve is the total amount, the dashed red curve the amount of females, and the dotted line is the proportion of females. The curve indicates that the largest number of biographees lived during the first half of the 20th century. The total curve appears smooth and does not show sudden changes due to historical events, e.g., the Second World War. The female percentage reaches a local maximum during the late 19th century and is growing constantly from 1950.

BiographySampo portal also allows one to look at the properties of the biographees, such as their average lifespan depicted in Fig. 4. The average life span for all biographees is 70.2 years. When comparing the male and female biographees, women on average live up to 72.2 years and men 69.8 years of age. Most biographees have died during their adulthood, but there are a few exceptions. For example, Sigfrid Jusélius (1887–1898)[33], who died at the age of 11, was included in the collection because her father, the well-

[32] http://biografiasampo.fi/tilastot/palkit

[33] https://biografiasampo.fi/henkilo/p4018

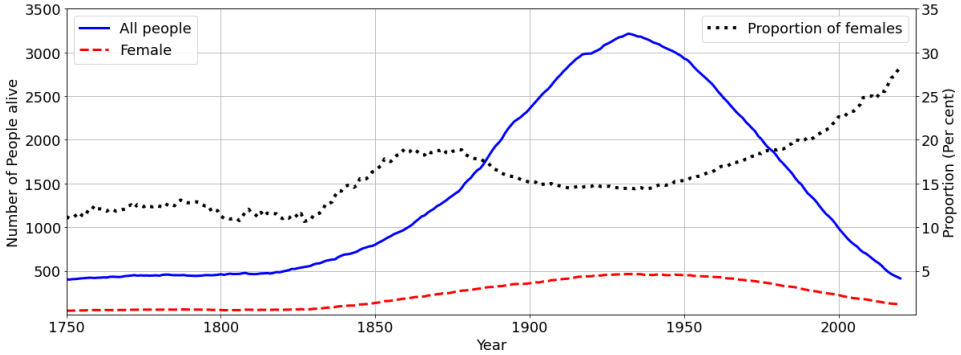


FIG. 3.: Number of male and female biographees alive on a timeline

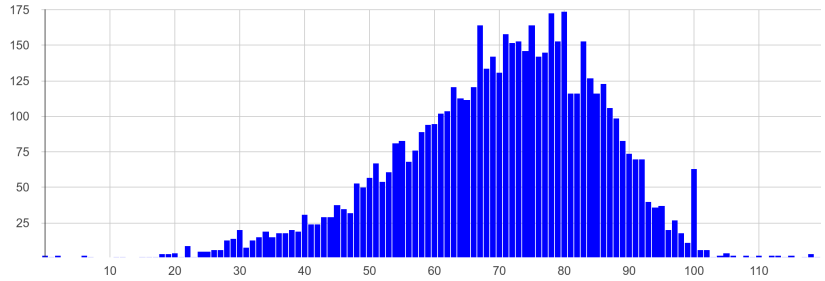


FIG. 4.: Average lifespan of the biographee's; screenshot from the BiographySampo portal

known tycoon Fritz Arthur Jusélius (1855–1930)[34] founded with his will the Sigfrid Jusélius Foundation[35] to promote medical research. Another example is soldier Yrjö Saarenpuu (1901–1919)[36] who was executed in a peculiar situation at the age of 19 instead of another person. There also seems to be quite a few biographees who lived 100 years old. However, the peek at 100 years is not a fact but results from the underlying data. At the moment, the underlying data does not tell whether a year, such as 1100 is rounded, or actually is a precise value.

The statistics application perspective of BiographySampo gives also insight into the life events of the biographies, such as getting married or having children. For example, Fig. 5 shows that the biographees got married on average at the age of 29 but there are also a few teen marriages and some older couples. A comparison of male and female biographees shows that women marry younger at the age of 26 than men at the age of 30 years. Men also marry more often after the age of 60 years.

There are also statistics about the number of children and spouses in the portal. The Fig. 7 represents the amount of children and the Fig. 6 the number of spouses for women and men. These plots are taken from the BiographySampo's statistics comparison view. Women's statistics are on the left hand side whereas the men's statistics are on the right hand side. Based on the statistics most women are married but have no children whereas men are mostly married to one partner and have no children. On average men have more children than women. Based on further data analysis using SPARQL queries [37], there are approximately 30.3% (286) of women and 9.32% (493) of men who are unmarried and childless. Using a different SPARQL query [38] it can be noted that the most common vocation for these childless and unmarried women is a teacher whereas for men it's a professor.

The BiographySampo portal allows users to generate statistical visualizations of correlations between, e.g., vocations or places of birth or death between biographees and their relatives. The Sankey diagram in

[34] https://biografiasampo.fi/henkilo/p4017
[35] https://www.sigridjuselius.fi/en/
[36] https://biografiasampo.fi/henkilo/p5253

[37] Query amount of unmarried and childless men and women: https://api.triplydb.com/s/oc6bZUcvp
[38] Query most common jobs for unmarried and childless persons: https://api.triplydb.com/s/Wtj8eUkhZ

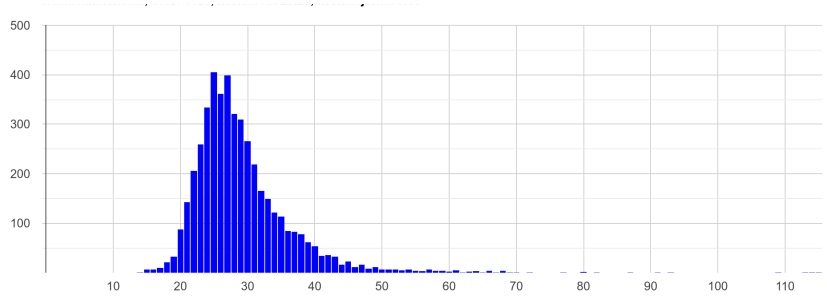
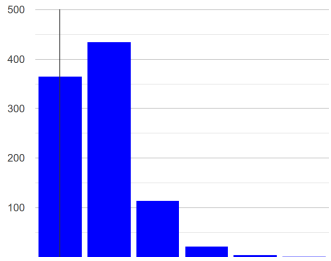
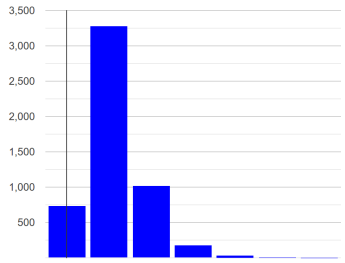


FIG. 5.: Average age of marriage; screenshot from the BiographySampo portal

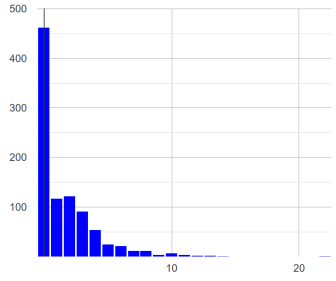


(A) Women

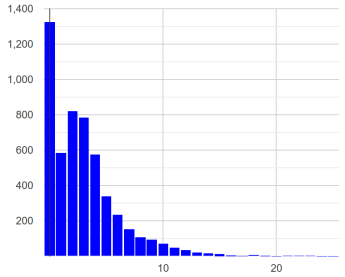


(B) Men

FIG. 6.: Average number of spouses for female and male biographees; screenshots from the BiographySampo portal



(A) Women



(B) Men

FIG. 7.: Average number of children for female and male biographees; screenshots from the BiographySampo portal

Fig. 8 visualizes correlations between the vocations of spouses so that husbands' vocations are on the left and their wives' on the right. The visualization suggests, for example, that men having a vocation related to theater often have an actress (*näyttelijä* in Finnish) as a wife. However, a wife of men of nobility gets a title of a baroness (*vapaaherratar* in Finnish). On the other hand, in cases like a farmer the vocation of a wife is not mentioned in the data at all.

3.1.1. Vocations

The NBF dataset also contains the vocations of each biographee except for 116 people. In this article the terms vocation and vocational group are used instead of terms occupation and occupational group. The vocation term is used because the person data contains in addition to occupational titles also, for example, honorary titles, academic degrees, and ranks of the peerage.

The biographees were distributed into vocational groups already at the stage when the collection was being mapped out by the editorial board. They chose to use a fairly standardized vocational classification previously used by other research projects in the 1980's, which was slightly modified to include all vocational groups in the NBF.

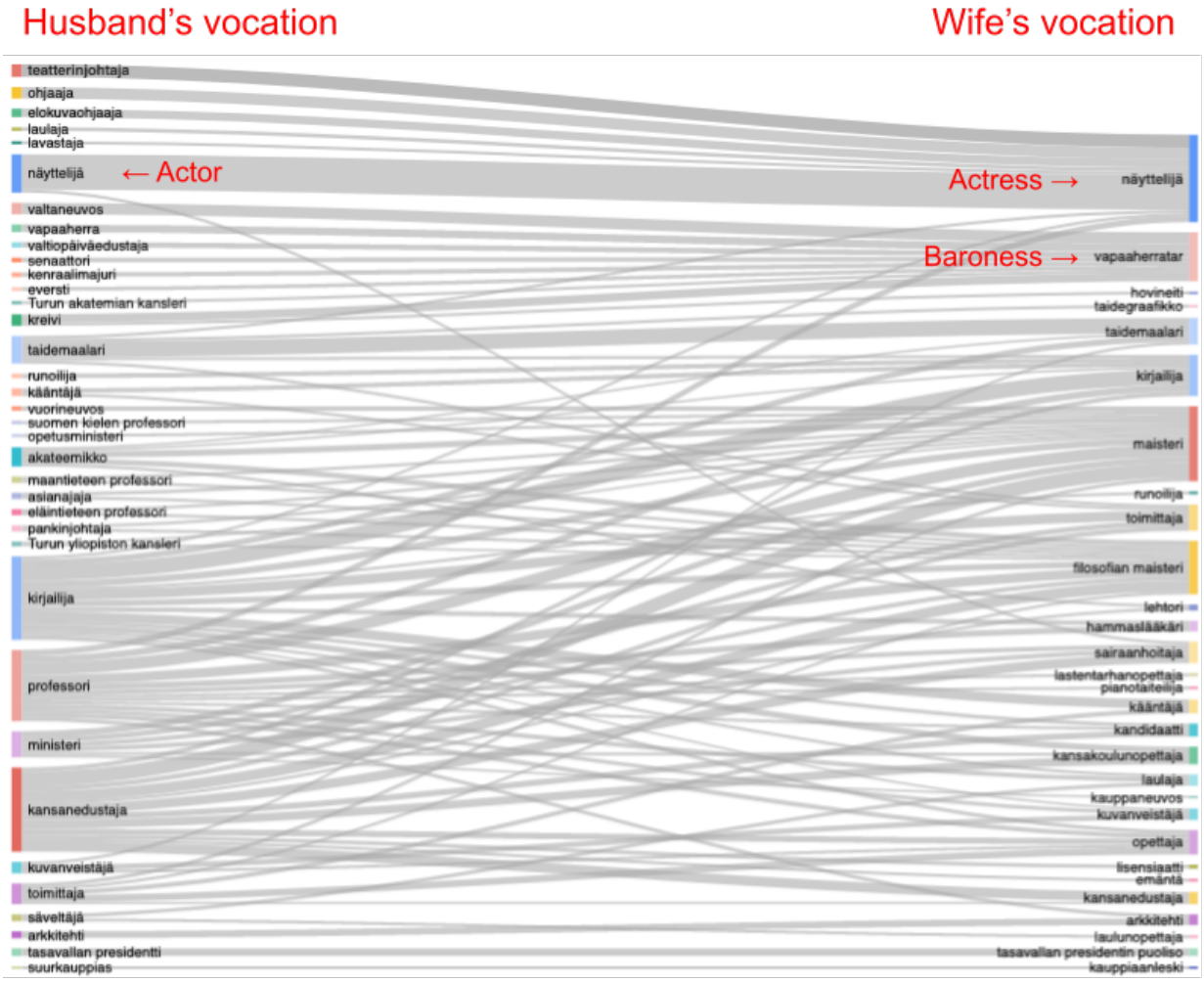


FIG. 8.: Sankey diagram depicting the correlations between the vocations of husbands and wifes; screenshot from the BiographySampo portal with English translations in red text

The use of vocational groups has a dual goal. On one hand they gave the editorial board a means to compose a diverse collection of biographies, and on the other hand they give the reader one more possibility to search the biographies. The vocational groups made it possible to take into account the different sectors and periods of Finnish history in selecting the biographees. The vocational groups are also useful as a search feature since they categorize the different titles (e.g., prime minister) to domains (e.g., politics).

Table 1 lists the 10 most common vocations for all, female and male biographees. The number in parentheses after the vocation indicates the number of occurrences. The list of the most common vocations for all and for men are similar but may have a different order of titles. The most common ones of these vocations appear for both female and male biographees. However, there are vocations which are more related to only one gender, like Lutheran minister and merchant for males, or actress and queen for females. The queen appears in the female vocations because the dataset contains all the historical rulers of Finland with their spouses.

In addition to vocations, there are also vocational groups for each biographee in the data. The vocational groups categorize the different titles, such as director, to different domains. Figure 9 depicts the distribution of the most common vocational groups in the NBF. In this figure, the vocational domains have been grouped based on the vocational grouping in the data. For example, musicians, authors, and artists are con-

TABLE 1: Most common vocations by gender

rank	Female	Male	All
1	Author (139)	Professor (1106)	Director (1182)
2	Director (125)	Director (1057)	Professor (1169)
3	Teacher (95)	Minister (443)	Author (501)
4	Professor (63)	Author (362)	Minister (481)
5	Painter (54)	Reporter (306)	Reporter (355)
6	Reporter (49)	Painter (203)	Painter (257)
7	Actress (46)	Lutheran minister (154)	Teacher (234)
8	Queen (45)	Merchant (144)	Scholar (159)
9	Unknown (40)	Scholar (140)	Merchant (158)
10	Minister (38)	Teacher (139)	Lutheran minister (154)

sidered to be in the group *Culture* whereas lawyers and judges are grouped to *Juridiciary*. However, many biographees have more than one vocation, and instead of selecting just one, they are all included in the visualization. The biographees have a maximum of 4 vocational groups and on average have 1.7 groups. For example, a person can be a judge and an author and is then included in both groups *Juridiciary* and *Culture*. The group *Charitable and NGO* consists of people working for charitable and non-governmental organizations (NGO) whereas *Other* contains marginal vocations, such as a member of the nobility, criminals, lovers, muses, fictional characters, and celebrities. The group *Unknown* is the proportion of biographees whose vocational group is unknown. The group of *Rewarded* is a heterogeneous group of people who have received a notable recognition for their work. This group was added into the list of vocational groups because it was a significant group of approximately 900 biographies. With all this in mind, based on the chart, the largest vocational groups within the NBF are *Culture*, *Politics*, *Science*, and *Economics*. From all the biographees, 50% of vocations belong to the four most popular groups. Similar visualization can be found from the ODNB [16] but vocational categories (areas of renown) differ.

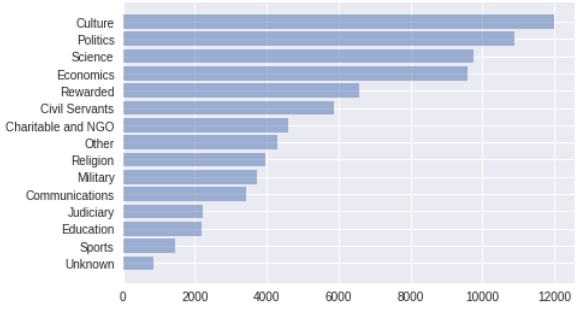


FIG. 9.: Most common vocational groups in the NBF

As mentioned earlier, a biographee can belong to more than one vocational group. The Fig. 10 depicts the most common intersecting vocational groups for a biographee who has more than one vocational group. For example, Field Marshal, president Gustaf Mannerheim (1867–1951)[39] was active in politics and in the military. In this diagram the diagonal consists of zeros because one biography cannot have one vocation more than once. When looking at the other vocational combinations, it can be seen that the people grouped into the group *Rewarded* are often also in the field of business and economic life or culture. Similarly, politicians are also often civil servants or working in economics. However, athletes have a very low correlation with the fields of science, religion, and the judiciary.

In addition to looking at the most common vocations and vocational groups, there is also a difference in most common vocations as a function of time which is depicted in Fig. 11 and 12. Figure 11 shows the ranking of 12 of the most common vocations and Fig. 12 the total amount of people with these vocations. The figures show that some vocations, e.g., director, professor, or author have a constantly high rank throughout the timeline. On the other hand, vocations like minister or reporter start gaining a higher rank during the late 19th century. Actor gains its highest rank in the years 1930–50 and naturally there are no movie actors before the cinema was invented and brought to Finland. Furthermore, some vocations such as merchant or Lutheran minister descend in the rank in the 19th century.

3.1.2. Relatives and vocations

The biographies have 5410 mentions of a father and 5310 mentions of a mother. In 619 cases the father also has a biographical entry, 94 of the mothers have biographies. Generally, especially with earlier biographees it is common that the vocation of a mother is not mentioned. There are approx. 5850 mothers whose vocation remains unknown, while 1130 fathers are missing this information. As an observation, there are, e.g., 340 cases where the father is a farmer, and 256 cases where he is a Lutheran minister. In cases like this, one could assume that the mother has been a farmer's wife, although it is not mentioned in the data entries.

Table 2 shows the 10 most common vocations of the biographees' parents. Six different columns where chosen similarly as in [16]. In the table teacher,

[39] http://biografiasampo.fi/henkilo/p328

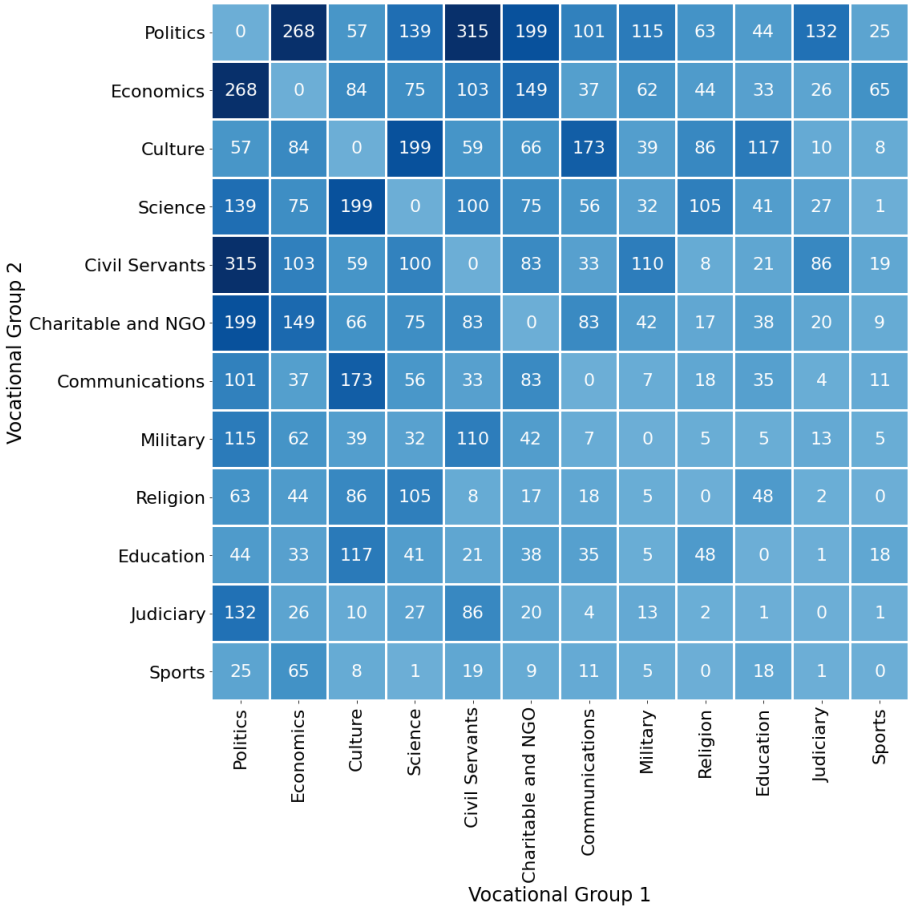


FIG. 10.: Correlations of the most common vocational groups

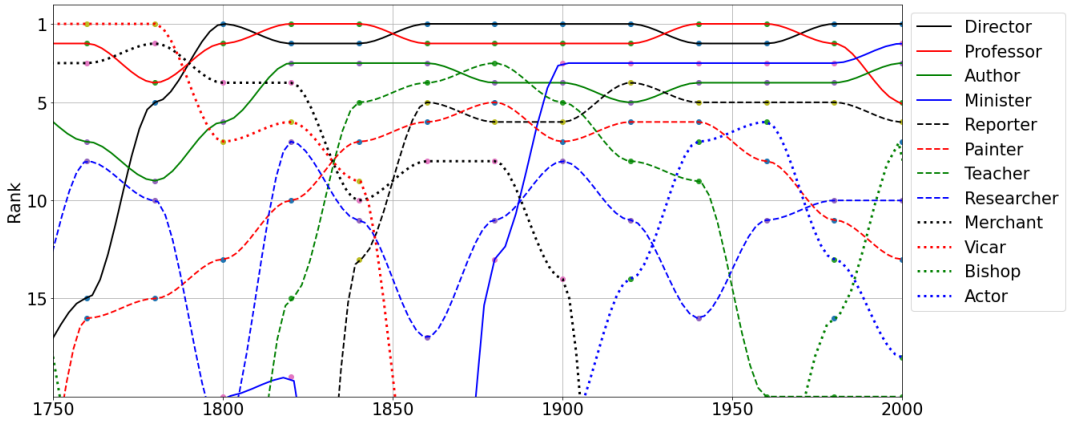


FIG. 11.: The most common vocations ranked on a timeline

farmer's wife, and nurse appear as the most common vocations of a mother, while farmer, director, and merchant as the most common of a father. On the other hand, some vocations of the biographees (Table 1) like minister, painter, or scholar do not appear in the parent data at all. Baroness and queen appear in the list of men's mothers, indicating that among nobility, the mother often has a biography entry in the dataset in

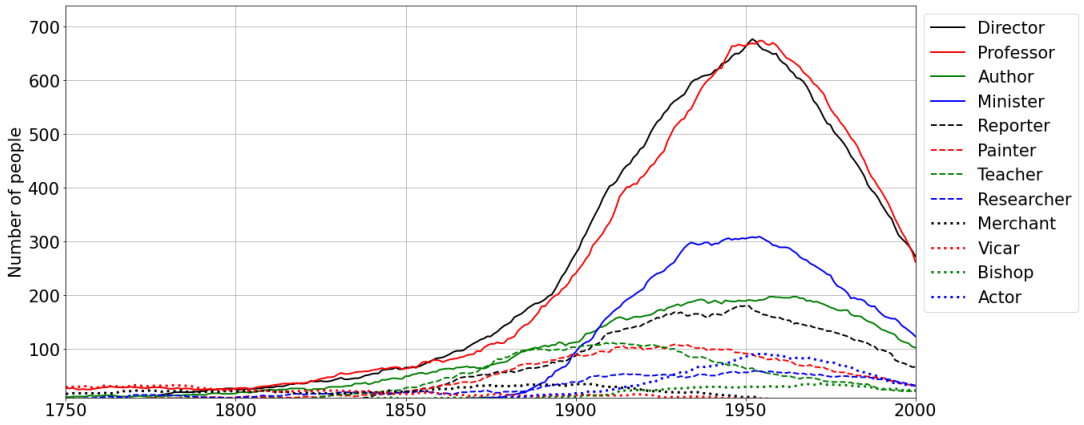


FIG. 12.: The most common vocations on a timeline

TABLE 2: Most common vocations of parents by gender.

rank	Women's Mothers	Men's Mothers	Women's Fathers	Men's Fathers	Women's Parents	Men's Parents
1	Teacher (23)	Teacher (89)	Farmer (52)	Farmer (378)	Director (57)	Farmer (380)
2	Farmer's wife (20)	Farmer's wife (59)	Director (51)	Merchant (250)	Farmer (53)	Merchant (263)
3	Nurse (9)	Nurse (25)	Merchant (44)	Director (236)	Merchant (44)	Director (245)
4	Seamstress (8)	Master of Art/Science ... (22)	Professor (35)	Lutheran minister (212)	Teacher (37)	Lutheran minister (214)
5	Director (6)	Baroness (21)	Lutheran minister (28)	Professor (161)	Professor (36)	Teacher (180)
6	Author (6)	Queen (16)	Proprietor (17)	Provost (124)	Lutheran minister (28)	Professor (164)
7	Master of Art/Science ... (5)	Lecturer (teacher) (14)	Provost (16)	Landed Peasant (113)	Farmer's wife (20)	Provost (124)
8	Actress (4)	Merchant (13)	Sea captain (14)	Teacher (91)	Proprietor (17)	Landed Peasant (123)
9	Servant (4)	Author (13)	Teacher (14)	Chaplain (88)	Reporter (17)	Chaplain (88)
10	Reporter (4)	Seamstress (12)	Blacksmith (13)	Blacksmith (83)	Nurse (16)	Blacksmith (83)
unknown	655	3910	225	1195	880	5105

her own right. The bottom row shows the number of cases where the information about a parent's vocation was not available.

Figure 13 depicts the correlation between the vocational groups of a child and his/her parents. The horizontal rows correspond to the groups of a child while the vertical columns to the groups of a parent. The number of biographees in each group is in the parenthesis after the group label. The values in the cells are normalized so that the values in each column sum up to one. To wit, the cell indicates the conditional probability for the group of child when the group of parent is known. Due to the dominant values at the diagonal of the matrix, there is an obvious correlation between the groups of a parent and of a child. The strongest correlations are found in the groups of *Culture*, *Politics*, and *Science*. Notice also how the off-diagonal values within the three groups are relatively low indicating a low intercorrelation and that they remain separated from each other. It can also be noticed that although *Agriculture* was a significant source of livelihood in Finland until the 1960's, the selection of biographies does not reflect that fact although many of the biographees came from farmer families.

3.2. *Events*

Events include the births and deaths converted from the structured CSV data, added with the lifetime events extracted from the semi-formal descriptions. An event usually contains a timespan and a possible reference to a place; we have extracted these mentions so that the event data can be illustrated on maps and timelines. The birth information was available for 6210 and death for 5800 out of the total of 6230 people.The semi-formal chapter of lifetime events was split into paragraphs describing the career, achievements (works, acknowledgments etc.), and a list of references. 5080 biographies contained a description of career and 3450 of achievements. Many of the people without the career description were historical figures of whom the records of education or vocations are not available. The data extraction generated 69 400 events of career, 29 900 events of achievement, and 18 000 mentions of honor.

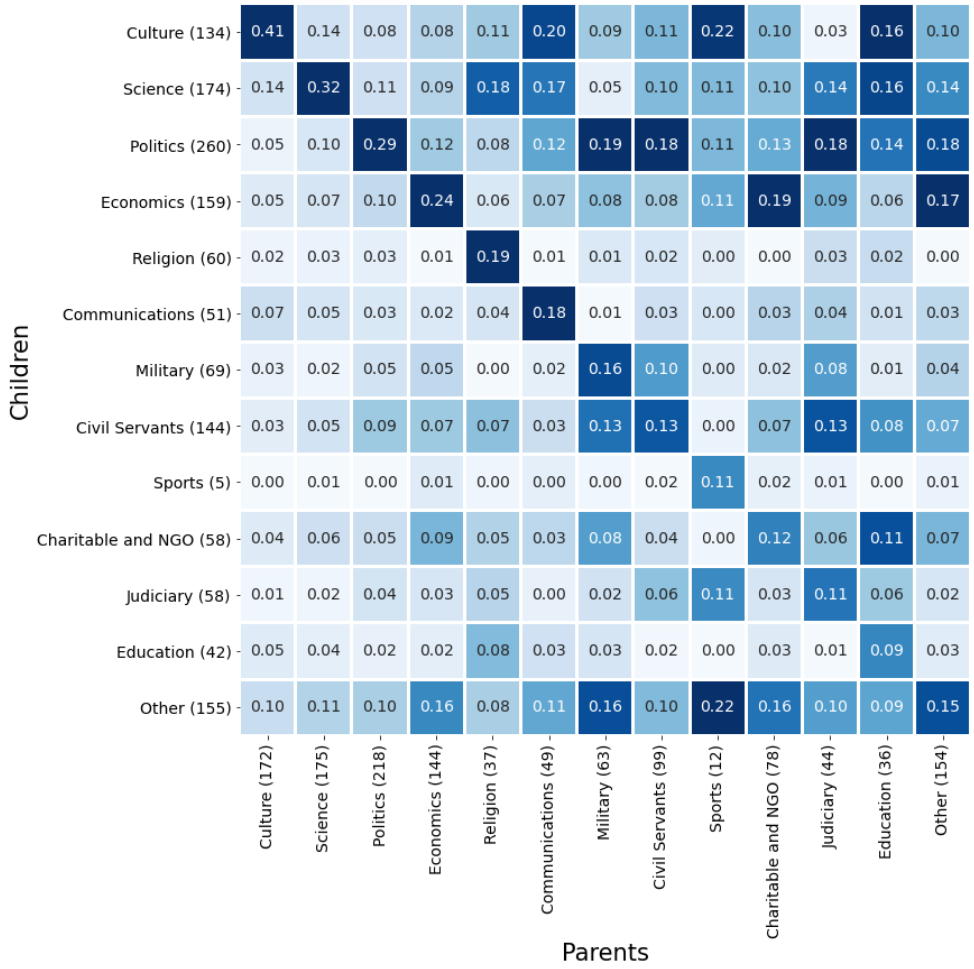


FIG. 13.: Correlations between the vocational groups of parents and children

FIG. 14.: Timeline with the number of events

The timeline in Fig. 14 depicts the number of events by year, e.g., births, deaths, and events related to a person's career. Generally the curve clearly follows the distribution of people alive shown in Fig. 3. The curve reaches the highest count around 1918, the time of the Russian revolution, of the beginning of Finland's independence and the Finnish Civil War. On the other hand, the curve shows a downwards peak in 1942, during the Second World War. This decrease is explained by the missing events in people's civil careers, although there

are military personnel in the people data. Furthermore, before the decade 1850 the data is so sparse and major events of that time, e.g., wars or plague pandemics, do not form distinct peaks to the figure.

3.3. Lives on Maps

Similarly to [16] we have ranked the ten most often mentioned places on a timeline in Fig. 15 but the illustration also contains names of towns and cities. The data was binned to intervals of 20 years. Helsinki became the capital of Finland in 1812 and has a constant highest ranking from the 1840's onward. The chart also shows a strong connection to Sweden with even more events than with the former capital Turku. Paris has had a high ranking during the latter half of the 19th century when it was a popular location for, e.g., university studies. The United States started to gain attention in the early 20th century. This attraction peaked during the decades 1940–1960. The old Finnish city of Vyborg lost its significance after the Second World War when it was annexed by the Soviet Union.

Figure 16 depicts a simplified illustration showing the referenced countries or continents. Generally biographees have had close connections to Sweden and Germany, and historically also to Russia, although it's significance has decreased during the 20th century. The Baltic Countries have increased their ranking after gaining independence from the Soviet Union. The third position of the United States after the 1940's is explained by, e.g., international studies. Africa has gained an increasing rank after 1960's due to, e.g., activities of development aid organized by the United Nations.

BiographySampo also provides the user with a map search view[40] in which the events extracted from the biographies are projected on the places where they occurred. After finding a place on the map, the place can be clicked. This opens a window showing the events with links to biographies. The maps in this view are not only contemporary ones but also historical maps served by the Finnish Ontology Service of Historical Places and Maps[41] [51], using a historical map service[42] based on Map Warper[43]. Many events of Finnish history took place in the eastern parts of the country that was annexed to the Soviet Union after the Second World War. Old Finnish places there may have been destroyed, place names have been changed, and names are now written in Russian. Using semi-transparent digitized historical maps on top of contemporary maps solves the problem by giving a better historical context for the events.

There is also a Life Maps application perspective in the portal. This perspective contains two kinds of prosopographical tools: 1) *Event maps* show how different events (births, deaths, career events, artistic creation events, and accolades) that a target group of people participated in are distributed on maps. 2) *Life charts* summarize the lives of persons from a transitional perspective as blue-red arrows from the birth places (blue end) to the places of death (red end). The prosopographical tools and visualizations in BiographySampo can be applied not only to one target group but also to two parallel groups in order to compare them. For example, Fig. 17 compares the life charts of male (on the left) and female (on the right) biographees in the NBF. This visualization suggests, perhaps surprisingly, higher international mobility of the female biographees. The arrows are interactive for close reading. For example, by clicking on the peculiar arrow to the north on the right, one sees that the feminist, activist and politician Annie Furuhjelm (1859–1937) was born in Alaska—Alaska and Finland both belonged to the Russian empire, and Annie Furuhjelms's father Hampus Furuhjelm was the governor of Alaska.

3.4. Reference Analysis and Networks

Based on the person data and extracted person references, the BiographySampo portal also contains network visualizations of people and how they are referenced in biographies. The networks enable the study of egocentric and socio-centric networks. In addition to using the BiographySampo portal, it is also possible to study the networks by using SPARQL queries to get the data. As an example, Fig. 18 depicts an extract around the vocational categories culture (marked with red) and politics (marked with blue) and black for other groups. The network is generated using the HTML links because of the coverage; currently the extracted person references are extracted for people born in the 1900s. HTML links referenced people in different datasets of SKS and were made only for the first occurrence of a biographee's name. The graph shows that the politicians form one solid cluster while the people who are grouped by their vocation to culture vocational group are divided into three smaller clusters, one rep-

[40] http://biografiasampo.fi/paikat/
[41] http://hipla.fi
[42] http://mapwarper.onki.fi
[43] https://github.com/timwaters/mapwarper

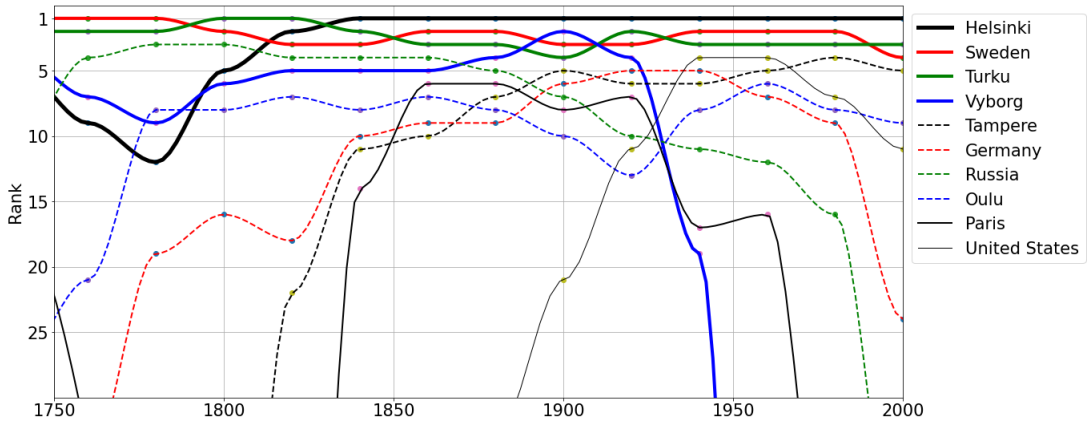


FIG. 15.: Top 10 places on a timeline

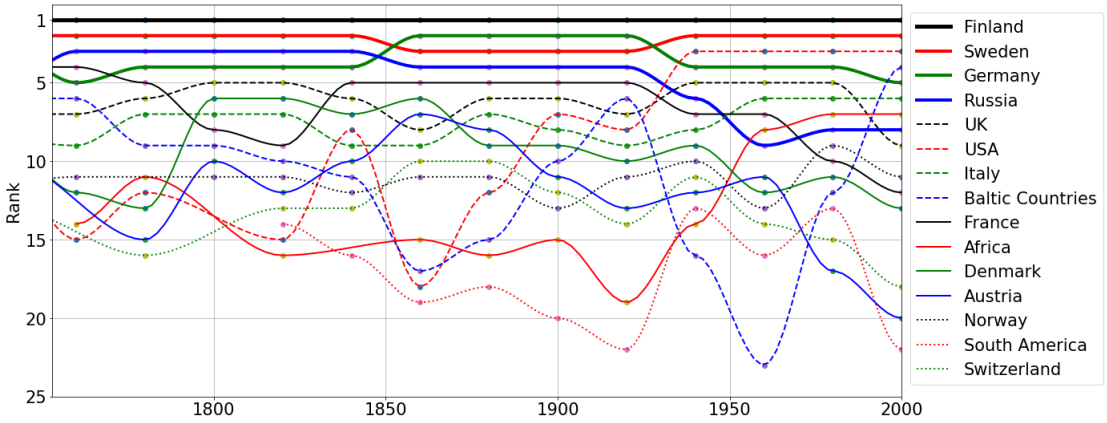


FIG. 16.: Top 15 countries on a timeline

resenting literature, one classical music, and one popular culture, when the corresponding biographies are analyzed by close reading.

3.4.1. Reference Analysis

In addition to enabling browsing of the data via networks, the tools in BiographySampo also enable link analysis currently only for biographies with HTML links. For each person, there is a view [44] where one can browse the references made to the biographee and to other biographies. The sentences containing the references are available from the linguistic RDF data and can be viewed in BiographySampo. For example, Fig. 19 shows the sentences that mention a) the biographee, here baroness Elisabeth Järnefelt (1839–1929)[45], in the other biographies, and b) the other biographees who are mentioned in her biography. These references show how a biographee is discussed in other biography texts, and how biographees are referenced in this biography. This is useful, for example, when studying the links in the egocentric networks. For example, in the egocentric network of the poet Aale Tynni (1913–1997)[46] there is a reference to the javelin thrower and film actor Tapio Rautavaara (1915–1979)[47], which seems odd. However, in this case the link analysis view explains the serendipitous connection: Aale Tynni and Tapio Rautavaara won gold medals in the 1948 Summer Olympics of London and they traveled together to receive their rewards.

BiographySampo also contains a chart for each biography, where the links from the source biography to other target biographies are calculated based on the birth decade of the target. This is illustrated in Fig. 20,

[44] http://biografiasampo.fi/henkilo/p3148/lauseet
[45] https://biografiasampo.fi/henkilo/p3148
[46] http://biografiasampo.fi/henkilo/p1238
[47] http://biografiasampo.fi/henkilo/p522

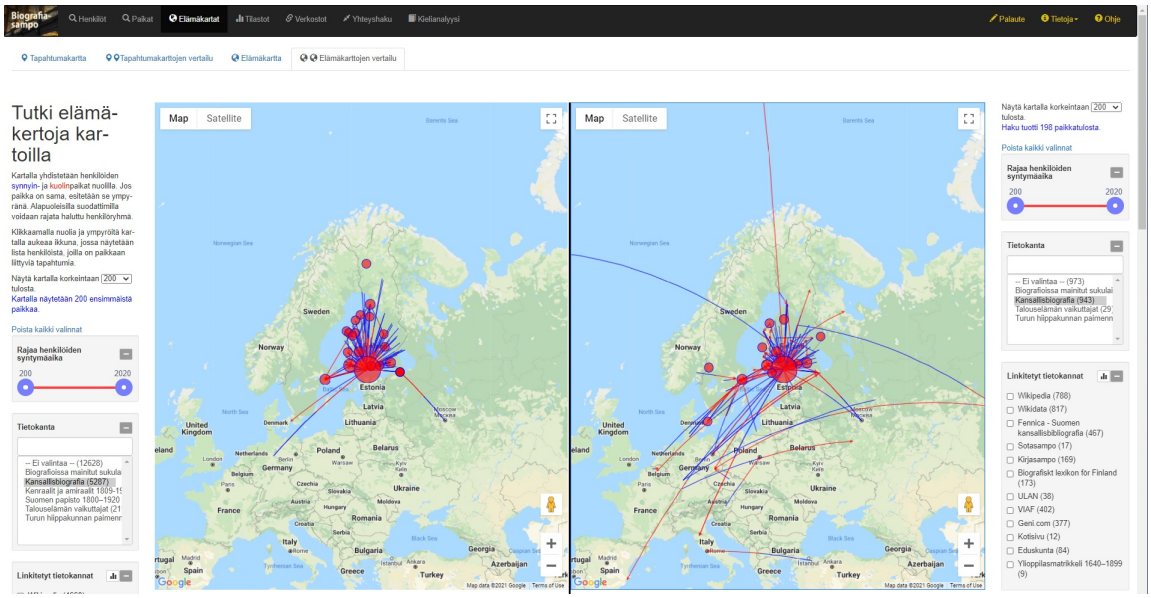


FIG. 17.: Comparing life maps of male (left) and female (right) biographees in the NBF in the BiographySampo portal

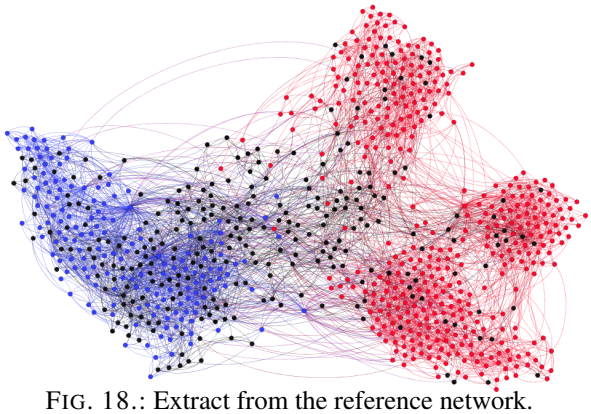


FIG. 18.: Extract from the reference network.

where the references of a source biographee and people referenced in the source's biography are plotted by their decade of birth. These plots show a) the influence of the source biographee by decade[48] and b) the prominent figures[49] mentioned in the biography of the biographee. This chart shows when the biographee influenced others the most or vice versa when people influencing the biographee were born. For example, a notable playwright can be mentioned frequently throughout history if the person's works are used by directors to recreate the scripts on stage or in movies.

In the BiographySampo portal there are no ready-to-use tools for counting references between biographies. In situations like this, one can use the data service SPARQL API directly to find out, for example, based on the HTML links who are the most often referred or "important" biographees. In Table 3 is the list of the top 10 people most commonly referred in the biographies of women. Whereas Table 4 is based on counting the references from the biographies of men. In addition to counting the references, the tables contain corresponding listings in the right column based on the PageRank measure of the reference network. The PageRank measure and algorithm [52, 53] was developed in Google to sort search results in a relevance order: the idea is to calculate the web pages' importance recursively based on the number of times the page is referred to and the PageRank of the referencing nodes, which emphasizes the value of references from highly ranked pages. Using the PageRank method leads to quite different ranking orders from the counting based rankings.

The PageRank measures have been calculated using the NetworkX Python library[50] after extracting the group of biographies from the SPARQL endpoint. A weighted network of biographies was created and was used for calculating the weight of the edges based on how many times there was a reference to a particular

[48] i.e. by the birth year of the person whose biography references the source biographee

[49] by their decade of birth

[50] https://networkx.github.io/

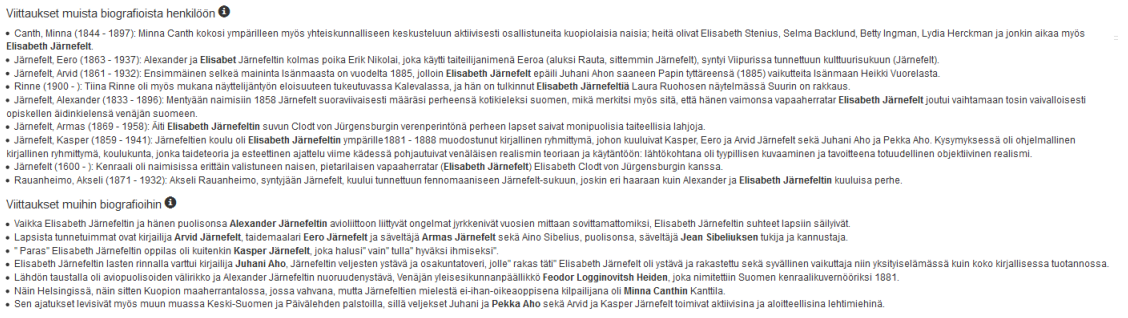


FIG. 19.: Sentences that reference people.

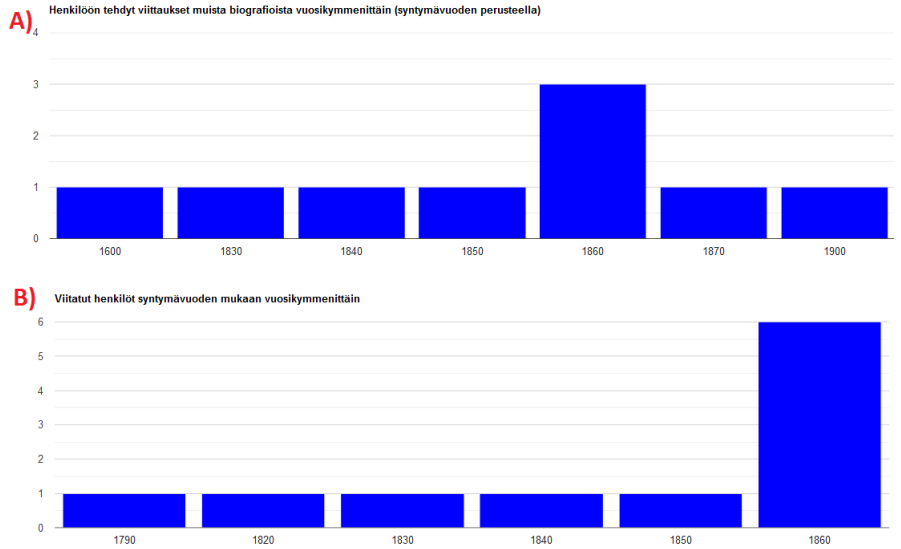


FIG. 20.: Plotting number of references by decade using the BiographySampo portal

TABLE 3: Top 10 referenced people in female biographies

	Count	PageRank
1	author Zachris Topelius (1818–1898)	author Zachris Topelius (1818–1898)
2	author Johan Ludvig Runeberg (1804–1877)	author Minna Canth (1844–1897)
3	president Urho Kekkonen (1900–1986)	singer Laila Kinnunen (1939–2000)
4	author Fredrika Runeberg (1807–1879)	politician Miina Sillanpää (1866–1952)
5	author Minna Canth (1844–1897)	author Fredrika Runeberg (1807–1879)
6	author Hilda Käkikoski (1864–1912)	author Marja-Liisa Vartio (1924–1966)
7	president Gustaf Mannerheim (1867–1951)	president Urho Kekkonen (1900–1986)
8	composer Jean Sibelius (1865–1957)	sculptor Essi Renvall (1911–1979)
9	painter Helene Schjerfbeck (1863–1946)	author Annikki Kariniemi (1913–1984)
10	painter Adolf von Becker (1831–1909)	painter Venny Soldan-Brofeldt (1863–1945)

biographee. The PageRank algorithm produces similar results to counting but the rank of a person changes. Women and therefore their networks are scarce causing the results between PageRank and counting the references to differ more. Women's list consists mainly of cultural influencers while men's have more politicians and rulers.

TABLE 4: Top 10 referenced people in male biographies

	Count	PageRank
1	president Gustaf Mannerheim (1867–1951)	president Urho Kekkonen (1900–1986)
2	president Urho Kekkonen (1900–1986)	president Gustaf Mannerheim (1867–1951)
3	president Juho Kusti Paasikivi (1870–1956)	king Gustav III of Sweden (1746–1792)
4	king Gustav III of Sweden (1746–1792)	president Juho Kusti Paasikivi (1870–1956)
5	author Johan Ludvig Runeberg (1804–1877)	author Johan Ludvig Runeberg (1804–1877)
6	author Zachris Topelius (1818–1898)	author Zachris Topelius (1818–1898)
7	prime minister Väinö Tanner (1881–1966)	king Charles XII of Sweden (1682–1718)
8	king Charles XII of Sweden (1682–1718)	prime minister Väinö Tanner (1881–1966)
9	composer Jean Sibelius (1865–1957)	composer Jean Sibelius (1865–1957)
10	president Kaarlo Juho Ståhlberg (1865–1952)	president Kaarlo Juho Ståhlberg (1865–1952)

Table 5 depicts the people with the highest centrality measures during chosen periods in the history of Finland. The data was generated by first generating the entire graph, and then filtering people related to each period and picking the ten people with the highest PageRank centrality measures [53]. The first column describes the years (–1809) when Finland was a part of Sweden. The first row under the header has the number of people during each period. Most of the people in the first column are monarchs of Russia or Sweden with Peter the Great, Emperor of Russian, on the first place and Empress Elizabeth on the second. Next, during the time in the second column (1809–1917) the Grand Duchy of Finland was an autonomous part of the Russian Empire. In contrast to the first column, the highly ranked people are not monarchs but prominent figures in Finnish culture and politics, such as the politician J.V. Snellman, and the poets and writers J. L. Runeberg and Z. Topelius. The third column covering the early years of the Finnish independence 1918–1944 contains mostly presidents and significant politicians of the era like the fourth column of years 1945–1994 between the Second War World and joining the European Union. One can, e.g., notice that presidents Paasikivi and Kekkonen as well as Field Marshal, president Mannerheim are present in both columns. In general, all the columns during the Finnish independence (1918–) are dominated by politicians.

3.4.2. References by Gender and between Relatives

Out of the references from male biographies 93.3% refer to a male biography, whereas only 6.7% refer to a female biography. On the other hand, from the female biographies 28.2% refer to a female biography. The average amount of links in a biography is 4.18 and there is no significant difference between the genders.

The difference between the ages of linked biographees was also studied with the observation that on average the mentioned person is 6.18 years older than the biographee. However, for females the average is 8.93 years while for men 5.73. A histogram of age differences is depicted in Fig. 21, where the negative values refer to an older person. The histogram shows that the modes of female and male distributions are both around zero, indicating that all people have plenty of links to people of nearly the same age. On the other hand, females have more links to people who are 20–75 years older while men have more links to people who are 10–50 years older than they. These statistics where calculated by picking random samples of the same size from both genders in order to avoid the male dominating bias in the data. This observation may be partly explained by the more frequent mentions of relatives in female biographies.

Table 6 shows the percentage of references between a biographee and his/her relative who is also a biographee. The studied relations are parents, spouses, children, siblings, and other relatives, e.g., cousins, grandparents and -children, or in-law-relatives. The table clearly indicates that females have in general more relatives in the dataset. Females have in average 2.11% of relatives mentioned in their biographies, while the corresponding value for men is 1.17%. Especially the spouse is mentioned in 0.74% of female biographies, while only in 0.11% of male biographies.

Figure 22 depicts the correlation between the vocational groups of two linked biographees. The numeric values of rows, columns, and cells follow the same principle as in Figure 13. The strongest correlations are found in the groups of *culture*, *politics*, and *science*. These three major dominant groups also appear as separated from each other due to their low corre-

TABLE 5:
People with highest PageRank values during five historical periods

	–1808	1809–1917	1918–1944	1945–1994	1995–
# of people	1270	2519	2682	2623	910
1	emperor Peter the Great	senator Johan V. Snellman	president Gustaf Mannerheim	president Urho Kekkonen	president Mauno Koivisto
2	empress Elizabeth of Russia	governor-general Nikolai I. Bobrikov	president Juho K. Paasikivi	president Juho K. Paasikivi	politician Jörn Donner
3	king Gustav III of Sweden	author Johan L. Runeberg	president Pehr E. Svinhufvud	prime minister Väinö Tanner	prime minister Paavo Lipponen
4	empress Catherine the Great	author Zachris Topelius	president Urho Kekkonen	president Mauno Koivisto	prime minister Kalevi Sorsa
5	emperor Peter III of Russia	professor Elias Lönnrot	president Kaarlo J. Ståhlberg	president Gustaf Mannerheim	politician Elisabeth Rehn
6	king Gustav I of Sweden	politician Georg Z. Yrjö-Koskinen	prime minister Väinö Tanner	attorney general Olavi Honka	president Tarja Halonen
7	king Charles IX of Sweden	politician Alexander Armfelt	composer Jean Sibelius	prime minister Karl-August Fagerholm	president Martti Ahtisaari
8	king Frederick I of Sweden	president Gustaf Mannerheim	prime minister Aimo K. Cajander	composer Jean Sibelius	prime minister Harri Holkeri
9	governor-general Per Brahe	emperor Nikolai I of Russia	president Kyösti Kallio	prime minister Vieno J. Sukselainen	politician Paavo Väyrynen
10	professor Henrik G. Porthan	statesman Arseni A. Zakrewsky	painter Akseli Gallen-Kallela	prime minister Rafael Paasio	author Bo Carpelan

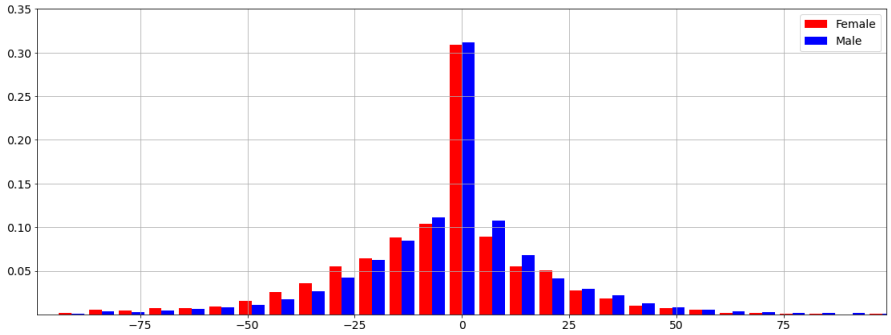


FIG. 21.: Histogram of differences in age of linked biographees

TABLE 6: Percentages of references to relatives by gender

	Parent	Spouse	Child	Sibling	Other older relative	Other younger relative	**Total**
Female	0.41	0.74	0.20	0.31	0.32	0.14	**2.11%**
Male	0.29	0.11	0.17	0.27	0.24	0.10	**1.17%**

lation. Groups like *religion* and *athletes* have plenty of references not only to these three major groups but also to themselves. On the other hand, these groups are rarely referenced from any other groups.

3.5. Network Metrics

The data has been enriched by linking mentions of people in the biographies, complementing the existing HTML links in the source data. The F-score of the HTML links in the source dataset is 97.3%. The result was calculated for 181 links from 35 biographies sampled randomly from the dataset. In few cases some biographies had not linked people who had a biography (mainly because they were written before the linking could be made), and in a couple cases the links pointed to wrong people. Some biographies had no links to other biographies. Typically, the biographies of athletes had no links because they only mentioned people such as team mates or coaches. The biographies are rarely written about coaches or lesser known athletes. In 75.5% of the biographies of athletes contained links while other vocational groups had links in over 81% of biographies, 88.2% of female and 89.8% of male biographees had links.The automatically extracted links add missing relations between biographees in addition to mentions of people who don't have biographies in the dataset. These automatically created links are used alongside the HTML links in the BiographySampo portal in a contextual reader application for the biographies and in reference networks[51].

Table 7 contains general metrics of the four networks, e.g., manually linked HTML network, automatically linked network, the network linked both manu-

[51] http://biografiasampo.fi/verkosto

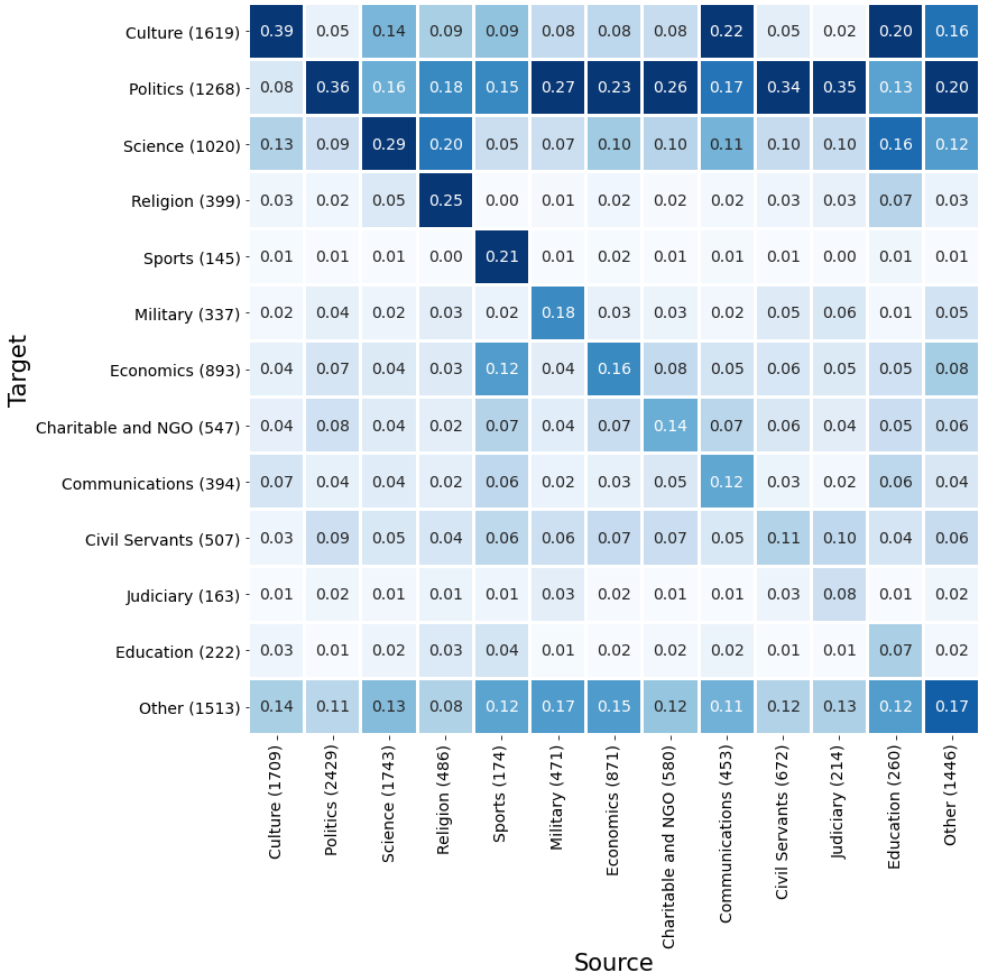


FIG. 22.: Correlations between the vocational groups of linked biographees

TABLE 7: Comparison between the four networks in the BiographySampo data using standard network metrics

	HTML links	Automatic	HTML + Automatic	Genealogical
nodes	5729	3247	5820	2487
edges	25013	12865	29464	3672
average degree	8.73	11.08	14.53	2.95
HD	430	557	986	19
max clique size	8	9	9	10
giant component	5664	3170	5779	428
number of components	30	35	20	585
diameter	11	12	11	30

ally and automatically, and the genealogical network. This table contains first the numbers of nodes and edges in the network. Average degree indicates the average amount of links for a single node and highest degree (HD) is the highest node degree in the network. Max clique size is the largest size of a clique, e.g., a value 8 indicates that there exists a subgroup of 8 people who all are linked to one another. The table shows

the number of separated components in the network, and the size of the largest connected component. It is to be observed that the genealogical network is scattered into numerous separated components, while the three reference networks are all more connected having giant components connecting most of the data points. The Diameter is the number of edges along the longest path between any two nodes in the network. Alpha (α) is the constant obtained when a power-law distribution is fitted on the degree distribution of the network. The Global Clustering Coefficient (CCG) is the measure of connected triples; the Average Path Length (APL) is the average number of edges traversed along the shortest paths for all possible pairs of the network nodes.

When comparing the results shown in Table 7 one has to remember how the Automatic references complete the graph of HTML links which is clearly shown by the measures of nodes and edge counts, average and highest degree, and giant component size. The last example network, the genealogical network is completely different by its nature where the people are linked by family relations.

Hashmi et al. [54] used a random sampling strategy for calculating the network measures in their study for structural similarity of social, communication, or collaboration networks. The example networks in their study are Twitter Friendship Network, Epinions Social Network, Wikipedia Vote Network, EU Email Communication Network, and Author Network. Their sampling strategy was to sample subgraphs of the size of 500 nodes with a breadth-first search and then calculate the values as average of ten such samples. Table 8 shows our reference networks in comparison with the five example networks analysed by Hashmi et al. where we used the same strategy to calculate the metrics. Comparing the values to their results shows that, e.g., the number of edges and therefore also the densities in our reference networks are in the same range as in Email and Author networks. Also the values indicating a small world or scale free behavior, e.g., CCG and α are in the same range as in the comparison networks. The smaller diameter in networks of BiographySampocan be explain by the degree distribution, approx. 75% of the nodes have a degree in the range 1 to 10.

3.6. Text Analysis

The biographies in BiographySampo can also be studied from a linguistic perspective in the Language Analysis view [52] of the portal. The Language view uses the linguistic knowledge graph to enable quantitative analysis of the biographical texts. Figure 23 shows in one of the plots in BiographySampo's Language View the average word count of biographies by decade. The histogram tells the typical length of biographies in different times based on the decade when the biographees were alive. This plot shows that the biographies of earlier people are somewhat shorter than the biographies concerning the 15th century, often due to the lack of data sources. However, when comparing this plot to the earlier distribution of the number of biographies by decade in Fig. 1, it can be seen that until the 19th century there are fewer biographies. This indicates that there may be a few longer biographies that distort the distribution of Fig. 23. For example, in the 16th century the biography of Mikael Agricola (1510–1557), a bishop who translated the New Testament into Finnish and developed Finnish into a written language, is several pages long whereas typical biographies of that time were only a page or two long, and in total there are approximately a little over 80 biographies. When looking at the number of biographies concerning the late 19th century, there are typically 500 biographies at the peak of the top decades.

In addition to the general statistics about the word count by decade, the user can get a list of the biographies with highest and lowest word counts. In Table 9, the top 10 of the longest and shortest biographies are listed based on their word counts. In the Table 9a of the longest biographies, the list mainly consists of politicians, presidents, and regents of Finland with one exception, Mikael Agricola. In Table 9b of the shortest biographies, there are people with different vocations, such as a local government official, two artists, a lesser known ruler, an athlete, and a priest. Most of the people in the list of the longest biographies are people who were in power or active during and after the World War II, such as president Urho Kekkonen. In the list of the shortest biographies, there are people who have been active in the Middle Ages or in the 18th and early 19th century.

In Table 10 the top 10 vocations that have the highest and lowest average word count in biographies are listed based on their word counts and on the number of biographies in the group. In Table 10a of vocations with the highest average word count, the list consists mainly of vocations that dominated also the list of bi-

[52] https://bit.ly/2PO8IVC

TABLE 8: Comparison between five example networks and reference networks of BiographySampo

	Twitter	Epinions	Wikipedia	Email	Author	HTML	Automatic	HTML + Automatic
edges	3099	13739	11672	2396	2404	2200	2678	2741
density	6.18	27.47	23.34	4.79	4.80	4.40	5.36	5.48
HD	237	278	281	499	102	159	403	323
diameter	11	7	12	7	10	5	5	5
CCG	0.19	0.43	0.35	0.54	0.60	0.36	0.34	0.35
APL	2.60	1.93	2.10	1.98	2.87	2.88	2.74	2.76
α	1.57	1.20	1.21	1.87	1.66	1.45	1.42	1.43

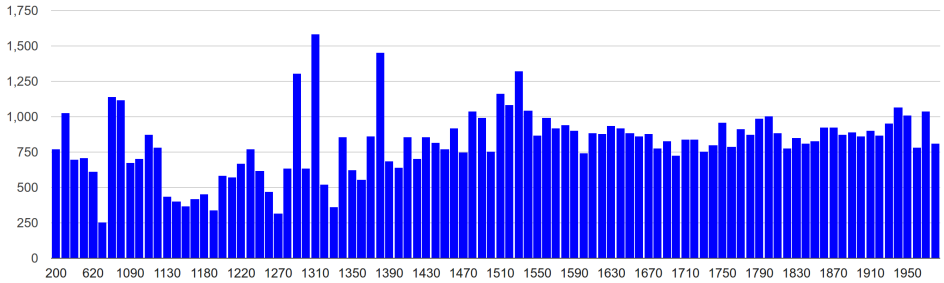


FIG. 23.: Amount of words in biographies by decade; screenshot from the BiographySampo portal

TABLE 9: Longest and shortest biographies

(A) Longest texts

Biography	Words
president Mauno Koivisto (1923–2017)	5369
president Gustaf Mannerheim (1867–1951)	4855
politician Otto Wille Kuusinen (1881–1964)	4717
senator Johan Vilhelm Snellman (1806–1881)	4656
prime minister Kalevi Sorsa (1930–2004)	4579
prime minister Edwin Linkomies (1894–1963)	4543
prime minister Rafael Paasio (1903–1980)	4462
bishop Mikael Agricola (1510–1557)	4171
queen Christina of Sweden (1626–1689)	4130
president Urho Kekkonen (1900–1986)	4075

(B) Shortest texts

Biography	Words
castle overseer Bengt Mårteninpoika (1442–1451)	174
lutheran minister Georg Stolpe (1778–1852)	174
bear hunter Per Huuskoinen (1732–1823)	174
lithographer Johan Henric Strömer (1807–1904)	177
painter Fridolf Weurlander (1851–1900)	177
writer Carl Fredrik von Burghausen (1811–1844)	180
king Kol of Sweden (?–1173)	197
mason master Petrus Murator de Kymitto (1466)	199
athlete Albin Stenroos (1889–1971)	201
demagogue Filippus (mentioned 1438)	205

ographees with the longest biographies by word count. The list's first group of the longest biographies has only 7 biographies by different authors and is about the lovers, muses, and favorites of politicians, artists, nobility, and military personnel who lived before the Finnish Independence. The other groups contain more biographies and have lower average word counts. In contrast, in the Table 9b lists the vocations with the shortest biographies (the lowest average word count). There are vocations, such as artisans, athletes, families, clergy, and government administrative officials. Some of these were found also on the list of the shortest biographies. The vocational group with the shortest biographies is athletes followed by artisans and judicial authorities.

In addition to word counts, the actual words and their frequencies can be listed for a filtered set of biographies. Table 11 lists the most common words (nouns, adjectives, and proper nouns) and the most common keywords for the whole NBF. The list of adjectives (Table 11c) contains common adjectives such as Finnish, new, first, great. These lists become more descriptive after the most common stop words are ignored. In the Table 11a, the most common keywords

TABLE 10: Top 10 longest and shortest texts by vocation

(A) Longest texts: average word count by vocation

Vocational group	Word count	Count
Favourites, muses, lovers	1377	7
Rulers and heads-of-state	1245	155
Administration (scientific communities)	1218	154
Theology	1088	87
Organizations, institutions	1081	30
Social sciences	1052	73
Politicians, activists	1049	308
Humanistic sciences	1048	396
Education and Cultural Work	1041	27
Nobility	1007	141

(B) Shortest texts: average word count by vocation

Vocational group	Word count	Count
Athletes	684	153
Artisans	696	80
Judicial authorities	702	264
Lawyers	728	59
Families	734	269
Local governments	746	151
Catholics	761	93
Agriculture and forestry	774	248
Regional administration	776	277
Trade, transport	786	384

are listed for the biographies and the number of times they appear (in column Count) in different biographies. The keywords have been extracted using the basic TF-IDF method from the nouns in the biographies. As can be seen from the table, this method typically picks up titles and other attributes related to the people described in the biographical texts, such as professors, kings, or women. In comparison, Table 11b lists the most common nouns in the biographies, containing similar words as in the keyword listing but in singular form (e.g., university and professor). However, these nouns constitute roughly 0.6% or less of the nouns and 0.2% or less of all the words in the dataset. All the keywords in the top 10 list can be found by looking at the top 50 nouns list.

As mentioned earlier, the user can select using facets any selection of the given data for inspection. As an example, we have selected the most common words used in the biographies of male and female politicians (e.g., MPs, presidents, ministers, rulers, and other political influencers in Finnish history). In Table 12 and Table 13 are the lists of the top ten nouns and adjectives for female and male politicians in BiographySampo. The table contains list of words for each group and the word count for the given word. Both lists have been created by querying from the biographical texts the top words of each part-of-speech group and filtering out most common words using a Finnish stop word list[53]. Both lists consist of mainly the same words but with some differences. In the female politician's list of nouns, the words for family life, such as spouse, son, daughter, and mother occur much more often whereas in the list of male politician's, nouns related to career, such as chairperson, post, and president are emphasized. The list of adjectives have similar words but with slight differences in order. However, when looking at lists of words that only exist in biographies of male or female politicians, for example, in lists of nouns and adjectives, the same themes are highlighted. Both groups have many terms that describe politics and career. But female politicians have a significant amount of nouns and adjectives that are related to family themes. Respectively, male politicians have a higher number of nouns and adjectives that describe economics, war, and religion.

[53] https://github.com/stopwords-iso/stopwords-fi

3.7. Author Analysis

In BiographySampo's dataset there are not only data about the biographees and their relatives but also about the authors of the biographical texts and their publishing dates. In this section statistics about the articles and their authors presented based on SPARQL queries to the data service.

The authors were chosen by the editorial board based on their expertise and previous research. Precedence was given to researchers who had recently published on the person in question or who had a deep knowledge of a specific field or period of history. The whole group of authors, more than 900 Finnish scholars, is so large and varied that it is very difficult to scrutinize them, especially because they come from so many fields of research. In addition to historians, they are specialists in various fields, e.g., art studies, jurisprudence, and medicine. The majority had a doctoral degree and a university affiliation. It is a group that can't be easily analyzed, since the information in the editorial database only includes their title and date of birth but not the affiliation or the field of study.

TABLE 11: Top 10 words and keywords in BiographySampo

(A) Top keywords

Keyword	English	Count
professorit	professors	536
kuninkaat	kings	427
yliopistot	universities	371
puolueet	political parties	370
teokset	works	312
naiset	women	283
sukulaiset	relatives	267
piispat	bishops	256
kirjailijat	writers	246
tutkimus	research	240

(B) Top nouns

Noun	English	Count
vuosi	year	30770
aika	time	19328
puheenjohtaja	chairman	12655
jäsen	member	11577
yliopisto	university	11391
lapsi	child	9709
professori	professor	8709
hallitus	government	8345
poika	boy	8216
historia	history	7250

(C) Top adjectives

Adjective	English	Count
suomalainen	Finnish	13381
uusi	new	11405
ensimmäinen	first	11344
suuri	great	10112
oma	own	8410
vanha	old	5939
nuori	young	5614
merkittävä	notable	4912
hyvä	good	4888
usea	several	4590

TABLE 12: Top ten words used in the biographies of female politicians

	NOUN			ADJ		
	Finnish	English	Count	Finnish	English	Count
1	nainen	woman	557	poliittinen	political	303
2	kuningatar	queen	459	vanha	old	169
3	puolue	political party	456	nuori	young	162
4	kuningas	king	422	seuraava	next	156
5	lapsi	child	378	suomalainen	Finnish	154
6	puoliso	spouse	317	yhteiskunnallinen	societal	122
7	eduskunta	parliament	314	merkittävä	significant	109
8	poika	son	283	sosiaalidemokraattinen	socialdemokratic	100
9	äiti	mother	283	tärkeä	important	97
10	puheenjohtaja	chairperson	278	kansainvälinen	international	94

TABLE 13: Top ten words used in the biographies of male politicians

	NOUN			ADJ		
	Finnish	English	Count	Finnish	English	Count
1	hallitus	government	4066	poliittinen	political	2493
2	puolue	political party	3766	suomalainen	Finnish	1453
3	tehtävä	task	2725	merkittävä	significant	1108
4	puheenjohtaja	chairperson	2649	tärkeä	important	1093
5	jäsen	member	2460	vanha	old	1078
6	kuningas	king	1845	keskeinen	central	995
7	toiminta	action	1840	nuori	young	985
8	eduskunta	parliament	1786	seuraava	next	983
9	sota	war	1742	sanottu	so called or said	693
10	presidentti	president	1718	yhteiskunnallinen	societal	646

The authors had to undertake to follow the guidelines and goals of the NBF, set by the editorial board. All articles were peer reviewed before being accepted for publication.

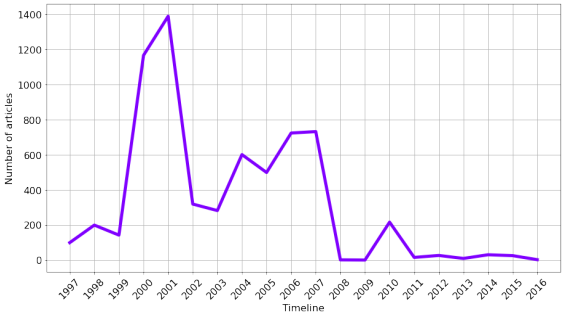


FIG. 24.: Number of articles written yearly in total

Since the publication of the NBF in print from 2003 to 2007, only 400 new biographies have been published. These newer articles were written thematically including biographies or people in different minorities, politicians, authors, actors and actresses, movie makers, theater directors, music educators, circus performers, and cartoonists.

The distribution of the number of articles published yearly can be seen in Fig. 24. The figure shows how the articles have been published from 1997 onward until 2016 (the most recent articles are not included in the BiographySampo). The figure has peaks before 2008 (the end of the publishing in print) and afterwards a minor peak in 2010 when a collection of new articles called the Multifaceted Finland was published online. Figure 25 depicts the distribution of how old the authors were when publishing biographies. The distribution also shows the difference between male and female authors.

Statistics about male and female authors of the biographies can be seen in Table 14, indicating also the gender of biographees they write about. The fraction of female writers is 32% of all writers in the dataset; the male writers dominate (68%) this dataset. There are three authors whose gender is unclear in the data, but they have written only 90 articles (approximately 1% of the articles). On closer inspection on whom the authors write about, it can be seen that men write mainly about men (94%) and women write about both genders. 41% of the female authors have so far written only about men and 26% about only women, while 5.7% of male authors write only about women.

Table 15 indicates that the female authors have written more often about people who are known influencers of culture, rewarded individuals, or people active in charitable or non-governmental organizations. In contrast to this, the male writers have mainly written about prominent politicians, scientists, or economical influencers. According to the editorial policies

TABLE 14: Breakdown of articles written by men and women

Gender	Women	Men
Writers	31.7%	68.0%
Articles	29.5%	69.1%
Write about women	39.1%	5.68%
Write about men	60.9%	94.3%
Only write about women	25.6%	4.52%
Only write about men	41.2%	79.5%
Write about both	33.2%	16.0%

of the NBF, the authors have not chosen their target biographees freely but were asked by the editors to write about particular people. The authors were selected based on what was known to be their areas of expertise.

4. Discussion

BiographySampo offers historians and the public data analytic tools that can be used for biographical and prosopographical research without experience in computer science by using the portal. With a little experience in formulating SPARQL queries and/or Python programming, the underlying SPARQL endpoint can be used for custom-made complex data analyses. In this paper, both approaches were used for creating historiographical analyses of the core part of the BiographySampo data, the National Biography of Finland. In addition, we have evaluated our methods to estimate the reliability of our results. Our approach gives scholars novel biographical and prosopographical tools for analyzing individual persons and their groups. The tools combine the quantitative approach and distant reading methods [55] with the qualitative approach, often based on close reading, typical to biographical research. The portal contains numerous views that enable the users to study the lives of the biographees as well as prosopographical groups in terms of statistics, maps, language usage, and networks based on references made in the biographies or based on the family relations extracted from the biographical descriptions.

The key findings of this paper give insight to the editors of the National Biography as well as to researchers in biography, prosopography, and historiography. They also highlight the possibilities and issues in modeling historical data related to, e.g, editorial choices, mod-

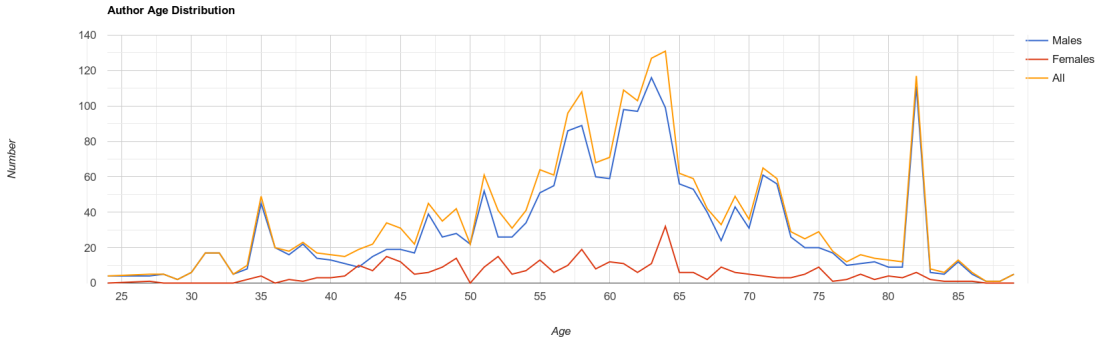


FIG. 25.: Author age distribution

TABLE 15: Most popular vocational groups of biographees for female and male authors

	Women			Men		
	Vocational group	Percentage	Count	Vocational group	Percentage	Count
1	Culture	42.6%	766	Politics	75.5%	1232
2	Politics	24.4%	398	Science	72.8%	1065
3	Economics	25.4%	365	Economics	73.3%	1053
4	Science	24.8%	363	Culture	54.1%	972
5	Rewarded	27.3%	269	Civil servants	81.7%	720
6	Charitable and NGO	27.3%	188	Rewarded	72.0%	710
7	Education	55.3%	183	Other	80.6%	518
8	Religion	28.3%	168	Military	90.0%	505
9	Civil servants	17.6%	155	Charitable and NGO	72.3%	498
10	Communications	23.8%	122	Religion	71.6%	425

eling uncertainty, serendipitous knowledge discovery, and data literacy.

Using automatically structured linked data in research needs new kind data literacy from the end user. As discussed above, in BiographySampo some parts (subgraphs) in the NBF dataset are based on reliable hand coded metadata while others were created by the machine. In big datasets like this it is not possible to check and correct the generated data manually, so more errors are expected to be encountered than in manually curated datasets. Furthermore, the linked data approach is based on using explicit classifications and ontologies for which different opinions may arise. In many cases, the underlying real world is too complex to be modelled fully in practice. For example, the historical place ontology underlying BiographySampo covers centuries of places that in reality change in time. For example, Finland was part of Sweden until 1809, then part of Russia until becoming independent in 1917, and after that some parts of her were annexed to the Soviet Union that became later the modern Russia.

The gaps in describing the lives of historical figures caused also challenges for analytics and data modeling. There are irregularities in describing biographees, their relatives, and vocations due to lack of reliable historical sources. This makes knowledge extraction somewhat challenging at times and the possibility for errors can increase, as the algorithms may misinterpret the original data and skip or mislabel data resulting in, for example, mislabeled family relations and anomalies in statistical or network visualizations. For example, similarly to what is mentioned by [55], the exact birth and death years of some people who lived in the early days of history are not known precisely, and heavily rounded inexact dates, such as 1100, appear in the data. The source data does not tell whether a year, such as 1100, is rounded or actually is a precise value. Without better knowledge, the system now assumes that all dates are accurate, resulting,e.g., in a peak of 100-year-old people in statistical visualizations. This phenomenon indicates how source criticism and understanding the underlying data is needed when interpreting quantitative results. A mechanism for rep-

resenting uncertainty in a machine understandable way would be needed to address the problem, but it remains a topic for future research.

In our work, the data was transformed from the CSV format to RDF and used as an input for further enrichment and transformation. Modelling the person and document metadata as RDF facilitated to creating the visualizations and performing the analyses depicted in this article. The transformation, extraction, and linking of the data was performed with satisfactory results (cf. Section 2.2). This data was used to enable distant reading by building data analytical applications and visualizations into BiographySampo. Unlike in [16, 17, 24], the data is in RDF format stored as knowledge graphs.

The Linked Data infrastructure created for BiographySampo also enables serendipitous knowledge discovery. The user can not only learn about the demographics through the statistical lens but also the connections between individual biographees through the network visualizations and reference analysis tools. The transformed knowledge graphs are published openly and can be queried with SPARQL to learn more about the data and the demographics.

Based on the analytics presented in this paper we have shown how to use Linked Data and SPARQL to create statistical, linguistic, and network analytics and visualizations to study a biographical data collection and its demographic features. These applications are related to the analytics represented in [16, 17, 24] but extend these analytics to describe the NBF dataset and also consider how the data has been created and used [**?**]. The data quality is not only impacted by its modeling and transformation process but also by its biases and sometimes historical uncertainty that exists in the source data. In comparison to the Ainm [17], the NBF is also biased towards the period from the mid 19th century onward whereas the ODNB [16] covers a wider span of time between the 16th century and current times.

Similarly to the Ainm and the ODNB, the visualizations tell the history of both the nation and of the collection itself. The place visualizations in this paper conform mainly to Finnish historical narratives that are tied to its neighbouring and European countries. Similar themes are present in the visualizations regarding relatives and vocations. The social structures are different in different countries, and cannot be used easily for transnational comparisons. As in Ainm and ODNB, the demographic of our dataset consists mainly of men while women are a minority. Furthermore, the networks are also influenced by the authors' decisions as each reference to another person is based on a choice. This has also become evident through the language analysis, as the lists of most common words in biographies of women contain more words to describe families than in the biographies of men. However, the language usage requires closer inspection to sort out the influence of the authors and it remains as a future work.

The Linked Data approach presented in this paper helps one to describe and analyze a biography collection with its strengths and weaknesses for further research, and to find out points of interest for close reading. The methods, results, and insights presented for the NBF can be utilized in DH research for other similar collections to learn more about the demographics of the collection itself, the underlying history, and to evaluate the reliability of the results.

Acknowledgments Thanks to Mikko Kivelä, Jouni Tuominen, and other members of the Semantic Computing Research Group (SeCo) for inspirational discussions related to network analyses and Linked Data services. We would also like to thank Werner Scheltjens and the anonymous reviewers for valuable feedback and comments of the earlier version of the article. Our research was part of the project Texts as Data Services (Severi)[54], funded mainly by Business Finland, and the EU project In/Tangible European Heritage – Visual Analysis, Curation and Communication (InTaVia)[55]. CSC – IT Center for Science has provided computational resources for our projects.

References

[1] T. Keith, *Changing conceptions of National Biography*, Cambridge University Press, 2005. doi:10.1017/cbo9780511497582.

[2] M. Klinge (ed.), *Suomen kansallisbiografia 1–10*, Suomalaisen Kirjallisuuden Seura, Helsinki, Finland, 2003–2007, p. 9519.

[3] F. Moretti and A. Piazza, Graphs, Maps, Trees: Abstract Models for a Literary History, *Modern Language Quarterly* **68**(1) (2007), 132–135. doi:10.1215/00267929-2006-032.

[4] F. Moretti, *Distant Reading*, Verso Books, 2013.

[5] E. Hyvönen, "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.*, CEUR Workshop Proceedings, vol. 2612, 2020, pp. 373–378. http://ceur-ws.org/Vol-2612/poster1.pdf.

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

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

[6] E. Hyvönen, P. Leskinen, M. Tamper, H. Rantala, E. Ikkala, J. Tuominen and K. Keravuori, BiographySampo – Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research, in: *The Semantic Web - 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2-6, 2019, Proceedings*, Vol. 11503 LNCS, Springer International Publishing, 2019, pp. 574–589. ISSN 16113349. ISBN 9783030213473. doi:10.1007/978-3-030-21348-0_37.

[7] B. Roberts, *Biographical Research*, Understanding social research, Open University Press, 2002.

[8] K. Verboven, M. Carlier and J. Dumolyn, A short manual to the art of prosopography, in: *Prosopography approaches and applications. A handbook*, Unit for Prosopographical Research (Linacre College), 2007, pp. 35–70, doi: http://dx.doi.org/1854/8212.

[9] H. Hakosalo, S. Jalagin, M. Junila and H. Kurvinen, *Historiallinen elämä - Biografia ja historiantutkimus*, Suomalaisen Kirjallisuuden Seura (SKS), Helsinki, 2014, pp. 1–342.

[10] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*, Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool, Palo Alto, CA, 2011. doi:10.2200/S00334ED1V01Y201102WBE001.

[11] E. Hyvönen, *Publishing and using cultural heritage linked data on the semantic web*, Morgan & Claypool, Palo Alto, CA, 2012. doi:https://doi.org/10.2200/S00452ED1V01Y201210WBE003.

[12] E. Hyvönen, Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery, *Semantic Web – Interoperability, Usability, Applicability* **11**(1) (2020), 187–193. doi:10.3233/SW-190386.

[13] L. Rietveld and R. Hoekstra, The YASGUI family of SPARQL clients, *Semantic Web – Interoperability, Usability, Applicability* **8**(3) (2017), 373–383. doi:10.3233/SW-150197.

[14] M. Koho, E. Heino and E. Hyvönen, SPARQL Faceter-Client-side Faceted Search Based on SPARQL., in: *Joint Proceedings of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop co-located with the 13th Extended Semantic Web Conference ESWC 2016, Heraklion, Crete, Greece, May 30, 2016*, CEUR Workshop Proceedings, 2016.

[15] E. Ikkala, E. Hyvönen, H. Rantala and M. Koho, Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces, *Semantic Web – Interoperability, Usability, Applicability* (2021).

[16] C.N. Warren, 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. doi:10.22148/16.028.

[17] Ú. Bhreathnach, C. Burke, J.M. Fhinn, G.Ó. Cleircín and B.Ó. Raghallaigh, A quantitative analysis of biographical data from Ainm, the Irish-language Biographical Database, 2019, presented at the 3rd Conference on Biographical Data in a Digital World (BD 2019). http://doras.dcu.ie/23774/1/Ainm%20BD%20FINAL.docx.pdf.

[18] A. Jatowt, D. Kawai and K. Tanaka, Time-focused analysis of connectivity and popularity of historical persons in Wikipedia, *International Journal on Digital Libraries* **20**(4) (2019), 287–305. doi:10.1007/s00799-018-0231-4.

[19] D. Metilli, V. Bartalesi and C. Meghini, A Wikidata-based tool for building and visualising narratives, *International Journal on Digital Libraries* **20**(4) (2019), 417–432. doi:10.1007/s00799-019-00266-3.

[20] E. Hyvönen, J. Tuominen, M. Alonen and E. Mäkelä, Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets, in: *The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, Springer-Verlag, 2014, pp. 226–230. doi:10.1007/978-3-319-11955-7_24.

[21] S. ter Braake, A. Fokkens, R. Sluijter, T. Declerck and E. Wandl-Vogt (eds), BD2015 Biographical Data in a Digital World 2015, CEUR Workshop Proceedings, Vol. 1399, 2015.

[22] A. Fokkens, S. ter Braake, R. Sluijter, P. Arthur and E. Wandl-Vogt (eds), BD-2017 Biographical Data in a Digital World 2017, CEUR Workshop Proceedings, Vol. 2119, 2017.

[23] R. Larson, Bringing Lives to Light: Biography in Context. Final Project Report, 2010, University of Berkeley. http://metadata.berkeley.edu/Biography_Final_Report.pdf.

[24] C. Warren, D. Shore, J. Otis, L. Wang, M. Finegold and C. Shalizi, Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks, *Digital Humanities Quarterly* **10** (2016), 1–16.

[25] A. Langmead, J. Otis, C. Warren, S. Weingart and L. Zilinski, Towards Interoperable Network Ontologies for the Digital Humanities, *International Journal of Humanities and Arts Computing* **10** (2016). doi:http://dx.doi.org/10.3366/ijhac.2016.0157.

[26] E. Hyvönen, M. Alonen, E. Ikkala and E. Mäkelä, Life Stories as Event-based Linked Data: Case Semantic National Biography, in: *Proceedings of the ISWC 2014 Posters & Demonstrations Track, a track within the 13th International Semantic Web Conference (ISWC 2014) Riva del Garda, Italy, October 21, 2014.*, Vol. 1272, CEUR Workshop Proceedings, Vol. 1272, 2014, pp. 1–4.

[27] E. Hyvönen, P. Leskinen, M. Tamper, J. Tuominen and K. Keravuori, 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*, Vol. 2084, CEUR Workshop Proceedings, 2018, pp. 372–385.

[28] E. Hyvönen, P. Leskinen, E. Heino, J. Tuominen and L. Sirola, Reassembling and Enriching the Life Stories in Printed Biographical Registers: Norssi High School Alumni on the Semantic Web, in: *Proceedings, Language, Technology and Knowledge (LDK 2017)*, Vol. 10318 LNAI, Springer, Cham, 2017, pp. 113–119. doi:10.1007/978-3-319-59888-8_9.

[29] G. Miyakita, P. Leskinen and E. Hyvönen, 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*, Vol. 11197 LNCS, Springer International Publishing, 2018, pp. 150–162. doi:10.1007/978-3-030-01765-1_18.

[30] A. Gangemi, V. Presutti, D.R. Recupero, A.G. Nuzzolese, F. Draicchio and M. Mongiovì, Semantic Web Machine Reading with FRED, *Semantic Web – Interoperability, Usability, Applicability* **8**(6) (2017), 873–893. doi:10.3233/sw-160240.

[31] M.C. Pattuelli, M. Miller, L. Lange and H.K. Thorsen, 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*, Alliance of Digital Humanities Organizations (ADHO), 2013, pp. 337–339.

[32] A. Fokkens, S. ter Braake, N. Ockeloen, P. Vossen, S. Legêne, G. Schreiber and V. de Boer, *BiographyNet: Extracting Relations Between People and Events*, in: *Europa baut auf Biographien*, New Academic Press, Berlin, Germany, 2017, pp. 193–224.

[33] M. Rospocher, M. van Erp, P. Vossen, A. Fokkens, I. Aldabe, G. Rigau, A. Soroa, T. Ploeger and T. Bogaard, Building event-centric knowledge graphs from news, *Web Semantics: Science, Services and Agents on the WWW* **37** (2016), 132–151. doi:10.2139/ssrn.3199233.

[34] M. Schlögl and K. Lejtovicz, A Prosopographical Information System (APIS)., in: *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 Linz, Austria, November 6-7, 2017.*, Vol. 2119, CEUR Workshop Proceedings, 2018.

[35] Á.Z. Bernád and M. Kaiser, The Biographical Formula: Types and Dimensions of Biographical Networks, in: *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 Linz, Austria, November 6-7, 2017.*, Vol. 2119, CEUR Workshop Proceedings, 2018.

[36] V. Gunter, S. Matthias and G. Vogeler, Data exchange in practice: Towards a prosopographical API (Preprint), in: *Proceedings of the Third Conference on Biographical Data in a Digital World (BD 2019), Varna, Bulgaria, September, 2019.*, 2019.

[37] E. Hyvönen, P. Leskinen, M. Tamper, H. Rantala, E. Ikkala, J. Tuominen and K. Keravuori, 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.*, 2019.

[38] Y. Wu, H. Sun and C. Yan, An event timeline extraction method based on news corpus, in: *2017 IEEE 2nd International Conference on Big Data Analysis*, IEEE, 2017, pp. 697–702. doi:10.1109/icbda.2017.8078725.

[39] E. Hyvönen and H. Rantala, Knowledge-based Relation Discovery in Cultural Heritage Knowledge Graphs, in: *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, Copenhagen, Denmark, March 5-8, 2019.*, CEUR Workshop Proceedings, 2019, pp. 230–239. http://www.ceur-ws.org/Vol-2364/.

[40] M. Tamper, P. Leskinen, K. Apajalahti and E. Hyvönen, Using Biographical Texts as Linked Data for Prosopographical Research and Applications, in: *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. 7th International Conference, EuroMed 2018, Nicosia, Cyprus*, Springer-Verlag, 2018, pp. 125–137. doi:10.1007/978-3-030-01762-0_11.

[41] M. Tamper, E. Hyvönen and P. Leskinen, Visualizing and Analyzing Networks of Named Entities in Biographical Dictionaries for Digital Humanities Research, in: *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICling 2019)*, Springer–Verlag, 2019, Accepted. https://seco.cs.aalto.fi/publications/2019/tamper-et-al-cicling-2019.pdf.

[42] P. Leskinen and E. Hyvönen, Extracting Genealogical Networks of Linked Data from Biographical Texts, in: *The Semantic Web: ESWC 2019 Satellite Events*, Springer–Verlag, 2019, pp. 121–125. doi:10.1007/978-3-030-32327-1_24.

[43] J.L. Martinez-Rodriguez, A. Hogan and I. Lopez-Arevalo, 'Information Extraction Meets the Semantic Web: A Survey, *Semantic Web – Interoperability, Usability, Applicability* **11**(2) (2020), 255–335. doi:10.3366/ijhac.2015.0140.

[44] M. Tamper, P. Leskinen, J. Tuominen and E. 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)*, CEUR Workshop Proceedings, 2020, pp. 3–14.

[45] J. Tuominen, E. Hyvönen and P. Leskinen, Bio CRM: A Data Model for Representing Biographical Data for Prosopographical Research, in: *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 Linz, Austria, November 6-7, 2017.*, Vol. 2119, CEUR Workshop Proceedings, 2018.

[46] S. Hellmann, J. Lehmann and S. Auer, NIF: An ontology-based and linked-data-aware NLP Interchange Format, 2012. http://scholar.google.com.au/scholar?q=nlp2rdf+hellman&btnG=&hl=en&as_sdt=0%2C5&as_ylo=2010#5.

[47] S. Hellmann, J. Lehmann and S. Auer, Towards an ontology for representing strings, 2012. http://svn.aksw.org/papers/2012/WWW_NIF/public/string_ontology.pdf.

[48] S. Hellmann, J. Lehmann, S. Auer and M. Brümmer, 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*, Springer Berlin Heidelberg, 2013, pp. 98–113. doi:10.1007/978-3-642-41338-4_7.

[49] C. Chiarcos and C. Fäth, 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*, Vol. 10318 LNAI, Springer, Cham, 2017, pp. 74–88.

[50] P. Leskinen, E. Hyvönen and J. Tuominen, Analyzing and Visualizing Prosopographical Linked Data Based on Biographies, in: *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 Linz, Austria, November 6-7, 2017.*, Vol. 2119, 2018, pp. 39–44.

[51] E. Ikkala, J. Tuominen and E. Hyvönen, Contextualizing Historical Places in a Gazetteer by Using Historical Maps and Linked Data, in: *Digital Humanities 2016, Krakow, abstracts*, 2016, pp. 573–577. https://dh2016.adho.org/abstracts/.

[52] S. Brin and L. Page, The anatomy of a large-scale hypertextual web search engine, *Computer Networks* **30** (1998), 107–117. doi:10.1016/s0169-7552(98)00110-x.

[53] M. Bianchini, M. Gori and F. Scarselli, Inside PageRank, *ACM Transactions on Internet Technology (TOIT)* **5**(1) (2005), 92–128. doi:10.1145/1052934.1052938.

[54] A. Hashmi, F. Zaidi, A. Sallaberry and T. Mehmood, Are all social networks structurally similar?, in: *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, IEEE, 2012, pp. 310–314. doi:10.1109/asonam.2012.59.

[55] S. Jänicke, G. Franzini, M.F. Cheema and G. Scheuermann, Visual Text Analysis in Digital Humanities, in: *Computer Graphics Forum*, Vol. 36, Wiley Online Library, 2017, pp. 226–250. doi:https://doi.org/10.1111/cgf.12873.

Publication VIII

Petri Leskinen and Eero Hyvönen. Linked Open Data Service about Historical Finnish Academic People in 1640–1899. In *DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, Riga, Latvia, Sanita Reinsone, Inguna Skadiņa, Anda Baklāne, Jānis Daugavietis (editors), pages 284–292, CEUR Workshop Proceedings, vol. 2612. October 2020. online https://ceur-ws.org/Vol-2612/short14.pdf.

Proceedings of Digital Humanities in Nordic Countries (DHN 2020), Riga, CEUR Workshop Proceedings, March, 2020. Accepted.

Linked Open Data Service about Historical Finnish Academic People in 1640–1899

Petri Leskinen[1][0000−0003−2327−6942] and
Eero Hyvönen[1,2][0000−0003−1695−5840]

[1] Aalto University, Semantic Computing Research Group (SeCo), Finland, and
[2] University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG), Finland
http://seco.cs.aalto.fi/projects/yo-matrikkelit, http://heldig.fi, first.last@aalto.fi

Abstract. The Finnish registries "Ylioppilasmatrikkeli" 1640–1852 and 1853–1899 contain detailed biographical data about virtually every academic person in Finland during the time period. This paper presents first results on transforming these registries into a Linked Open Data service using the FAIR principles. The data is based on the student registries of the University of Helsinki, formerly the Royal Academy of Turku, that have been digitized, transliterated, and enriched with additional data about the people from various other registries. Our goal is to transform this largely textual data into Linked Open Data using named entity recognition and linking techniques, and to enrich the data further based on links to internal and external data sources and by reasoning new associations in the data. The data will be published as a Linked Open Data service on top of which tools for searching, browsing, and analyzing the data in biographical and prosopographical research are provided.

1 Biographical Dictionaries on the Web

Biographical dictionaries [18] have been published traditionally as printed book series. In 2004, the Oxford Dictionary of National Biography[3] (ODNB) was published on-line, and many major biographical dictionaries started to open their editions on the Web with search engines for finding and (close) reading biographies of interest.[4]

ODNB and other early adopters of web technology started the paradigm shift in publishing and reading biographical dictionaries on the Web. Related to our work on *BiographySampo – Finnish biographies on the Semantic web* [15] we have proposed that the next paradigm shift is to publish and use biographical dictionaries as Linked Data on the Semantic Web. [16] This paper presents first results of a new case study where this idea is applied to a new dataset: the

[3] `https://www.oxforddnb.com`

[4] On-line national biographical collections include, e.g., USA's American National Biography [1], Germany's Neue Deutsche Biographie [4], France's Nouvelle Biographie générale [5], Biography Portal of the Netherlands [2], Dictionary of Swedish National Biography [3], and National Biography of Finland[5] (NBF).

Finnish registries "Ylioppilasmatrikkeli" 1640–1899[6] that contain short biographical descriptions of 28 000 students of the University of Helsinki[7], originally the Royal Academy of Turku[8] in Finland. This publication covers a significant part of the history of Finland and the Finnish university institution, since the University of Helsinki was the only university in the country during the time frame in focus.

This paper presents an overview of research underway, addressing the problem of transforming biographical registers into Linked Data, and enriching their contents using Named Entity Recognition and Linking and by reasoning. The focus of this paper is on the data transformation process; application of the data in Digital Humanities research will be reported later. We first present the source dataset and the ontology model used for representing the biographical data in a semantic, i.e., machine "understandable" way. After this the underlying knowledge graph is discussed and its publication using the Linked Data Finland platform [17]. In conclusion, related works are discussed and relations of the work in a larger setting are summarized.

2 Source Datasets

An example of a registry entry for *Anders Israel Cajander*[9] is depicted in Fig. 1. The description starts with date or year of enrollment, in this case *11.2.1830*. After that there is the full name followed by the place and time of birth. Next there is a Finnish abbreviation *Vht* meaning parents; in the example case the father is *Zachris Johan Cajander* and the mother *Gustava Karolina Neiglick*. After that there are two lists of events; the events mentioned before the em dash (—) are related studies and academic career with the University of Helsinki; for example *Ylioppilas Helsingissä 11.2.1830.* (A student in Helsinki 11.2.1830). After the em dash there is another list of events during his career; e.g. *Äyräpään tuomiokunnan tuomari 1857* (Judge at the District Court of Äyräpää 1857). A person's death is marked with the symbol † and burial with ‡; the person in example died in Wyborg on December 18th 1901.

After the life time description there is a possible field for relatives; in the example case his spouse is mentioned first *Pso: 1841 Fredrika Emelie Schildt* where *Pso* is a Finnish abbreviation for *puoliso* (spouse). There are three relatives who also have an entry in the register, e.g. two brothers *Veli: Gustaf Adolf Cajander* and *Veli: Zakarias Cajander*, and a brother-in-law *Lanko: Berndt Vilhelm Kristoffer Schildt*. The author of the 1640–1852 dataset Yrjö Kotivuori has manually added links from a person's description to those relatives also found in the register, like the three relatives in the example case. Finally, at the

[6] The registry contains two parts: the database covering the years 1640–1852 is available in Finnish and Swedish at `https://ylioppilasmatrikkeli.helsinki.fi`, and the registry of 1853–1899 is at `https://ylioppilasmatrikkeli.helsinki.fi/1853-1899`

[7] `https://en.wikipedia.org/wiki/University_of_Helsinki`

[8] `https://en.wikipedia.org/wiki/Royal_Academy_of_Turku`

[9] `https://ylioppilasmatrikkeli.helsinki.fi/henkilo.php?id=14689`

end of the description there is a field for reference material used for collecting information about the person.

When designing the ontology model we wanted it to provide answers to possible research questions made by historians, such as: "Are there marriages between second cousins?", "Are there families that are closely connected by marriages?", and "What is the distance between the place of birth and the most long-term place of living and what are the mean and the standard deviation of this measure?".

11.2.1830 **Anders Israel Cajander** 14689. * Leppävirralla 24.2.1811. Vht: Savon alisen kihlakunnan kruununvouti *Zachris Johan Cajander* († 1862) ja *Gustava Karolina Neiglick*. Kuopion triviaalikoulun oppilas 4.2.1822 – 22.6.1826 (betyg). Viipurin lukion oppilas 17.9.1827 – 1.7.1829. Ylioppilas Helsingissä 11.2.1830 (arvosana approbatur cum laude äänimäärällä 14). Viipurilaisen osakunnan jäsen 12.2.1830 *12/2 1830 \ Anders Israel Cajander \ 24/2 1811 \ KronoFogden Zachr. Joh. Cajander i Randasalmi \ Leppävirta \ [med betyg] fr. Gymn. i Wiborg \ Uttog betyg d. 12/10 1833 för att ingå vid Rättegångsverken.* Merkitty oikeustieteellisen tiedekunnan nimikirjaan 9.10.1832. Savokarjalaisen osakunnan perustajajäsen 1833 *Anders Israël Cajander.* Tuomarintutkinto 10.12.1833. Vaasan hovioikeuden auskultantti 24.12.1833. — Varatuomari 1837. Kihlakunnantuomarin arvonimi 1847. Äyräpään tuomiokunnan tuomari 1857, Jääsken tuomiokunnan 1870, Rannan tuomiokunnan 1877, ero 1891. Hovioikeudenasessorin arvonimi 1868. Laamannin arvonimi 1870. Valtiopäivämies 1872. † Viipurissa 18.12.1901.

Pso: 1841 *Fredrika Emelie Schildt* († 1892).
Veli: Räisälän kappalainen *Gustaf Adolf Cajander* 15376 (yo 1835, † 1882).
Veli: kirjailija *Zakarias Cajander* 16147 (yo 1843, † 1895).
Lanko: lääninmetsänhoitajan apulainen *Berndt Vilhelm Kristoffer Schildt* 14968 (yo 1832, † 1892).

Viittauksia: HYK ms., Savokarj. osak. matr. #22; HYK ms., Viip. osak. matr. III #2037; HYKA, Album 1817–65 s. 230; HYKA OTA Ba, Oikeustieteellisen tiedekunnan nimikirja 1828–72 s. 19; KA Ansioluettelokokoelma. — T. Carpelan, Studentmatrikel (1928–30) s. 12; Matrikel öfver ungdomen vid Kuopio Trivialskola [1816–42]. Aarni 10 (1958) #572; H. Hornborg och I. Lundén Cronström, Viborgs gymnasium 1805–1842. Biografisk matrikel. SSLS 388 (1961) #311. — K. F. J. Schauman, Finlands jurister (1879) #35; H. J. Boström, Wasa Hofrätts auskultanter 1776–1876. SSV 5 (1921) s. 94–133 #293; H. Holmberg, Suomen tuomiokunnat ja kihlakunnantuomarit (1959) s. 236.

Fig. 1. Register entry for *Anders Israel Cajander*

3 Ontology Model for Representing Biographical Data

In addition to basic data, such as people's names and dates and places of birth and death, the source data provides rich content of information like the relatives, student nation, academic degrees, career events, and sources of reference. In our case we selected the data harmonization approach and the event-centric CIDOC CRM [8] ISO standard as the ontological basis, since biographies are based on life events. Bio CRM [26] is a domain specific extension of CIDOC CRM, applicable to biographical data; it extends CIDOC CRM by introducing role-centric modeling. Bio CRM has been used in our earlier projects of Norssi High School Alumni [14,23] and BiographySampo [16,22] to model roles and occupations as well as the relationships between people.

The ontology schema is depicted in Fig. 2. The people in the register are represented as instances of the class `foaf:Person` and the mentioned relatives using `:ReferencedPerson`. The resources of actor classes are enriched with lifetime events and relationships. Events, e.g. birth, baptism, enrollment, death, and burial, are subclasses of `:Event` and enriched with linking to corresponding times, places, and titles. The source data provides two kind of binary relationships: family relations (such as *parents, children, spouses, ...*) and domain-specific relations (such as *student, teacher, ...*). As an example of converted RDF, the data of Fig. 1 is depicted in Fig. 3 in Turtle format.

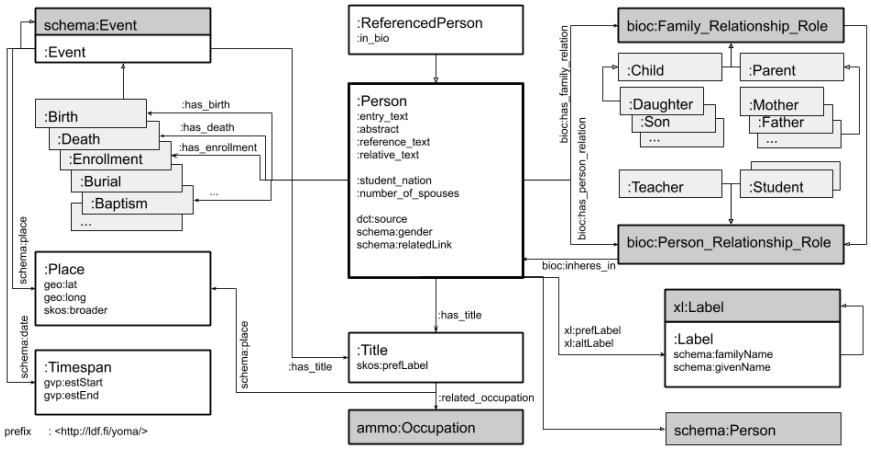


Fig. 2. Ontology schema for representing biographical data

4 Knowledge Graph of Historical Academic Persons

The extracted knowledge graph contains currently 28 000 students and 56 700 other people mentioned in the descriptions. These person resources are further

```
@prefix dct:   <http://purl.org/dc/terms/> .
@prefix foaf:  <http://xmlns.com/foaf/0.1/> .
@prefix schema: <http://schema.org/> .
@prefix skos:  <http://www.w3.org/2004/02/skos/core#> .
@prefix xl: <http://www.w3.org/2008/05/skos-xl#> .

@prefix :      <http://ldf.fi/yoma/> .
@prefix bioc:  <http://ldf.fi/schema/bioc/> .
@prefix event: <http://ldf.fi/yoma/event/> .
@prefix label: <http://ldf.fi/yoma/label/> .
@prefix rels:  <http://ldf.fi/yoma/relations/> .
@prefix titles: <http://ldf.fi/yoma/titles/> .

:p14689  a                  foaf:Person ;
        bioc:has_family_relation
                            rels:r2590153968717837790,
                            rels:r3067073318077691085,
                            ... rels:r2556529631795161483 ;
        :abstract           "<strong>Cajander, Anders Israel
                             </strong>, laamanni (yo 1830,
                             † 1901)"@fi ;
        :enrollment_text    "11.2.1830" ;
        :entry_text         "11.2.1830 <strong>Anders Israel
                            Cajander</strong> <a href= ...
                             ... († 1892).</p>\\"@fi ;
        :has_birth          event:b14689 ;
        :has_death          event:d14689 ;
        :has_enrollment     event:e1961333836730594986 ;
        :has_title          titles:v7140446880754877544 ;
        :id                 "14689" ;
        :number_of_spouses  1 ;
        :reference_text     "HYK ms., Savokarj. osak. matr. #22;
                             HYK ms., Viip. osak. matr. III #2037;
                             ... (1959) s. 236."@fi ;
        :relative_text      "<p>Veli: Räisälän kappalainen <em>
                             Gustaf Adolf Cajander</em>
                             ...  (yo 1832, † 1892).</p>"@fi ;
        :title_text         "laamanni" ;
        dct:source          :yo1640_1852 ;
        schema:gender       schema:Male ;
        schema:relatedLink
        <://ylioppilasmatrikkeli.helsinki.fi/henkilo.php?id=14689> ;
        skos:prefLabel      "Cajander, Anders Israel (1811-1901)"@fi ;
        xl:prefLabel        label:l2728541252431989123 .
```

Fig. 3. RDF/Turtle data for *Anders Israel Cajander*

enriched with 76 100 life time events and interlinked by 83 400 family and 3760 academic relations. There are 10 600 distinct occupational titles often referring to a place and an occupation, e.g., *the Bishop of Porvoo* or *Diving Commissioner who lived in Espoo*.

This information was extracted from description texts, which are structured with symbols (like † or ‡) and keywords (like *Vht* for parents or *Pso* for spouse) that help recognizing the content of each particular text field. To process the data, regular expressions, vocabularies of Finnish names, and a Python script recognizing different expressions of dates and times are used. Vocabularies of Finnish names are also used to infer a person's gender, when it is not otherwise obvious; generally, the person data is strongly male dominated, and the first female student entered the University of Helsinki in the year 1870[10].

The data set contains an ontology of more than 100 family relations. In percentage terms, most of the mentions are close relatives like *father* and *brother*, but occasionally there are, e.g., in-law-relations like *stepfather-in-law*, relations marked as potential with a question mark *son-in-law(?)*, or relations reaching over several generations like *uncle of the great great grandfather*. We extended our earlier ontology [23] to cover at least 99% of mentioned relationships. The data of the years 1640–1852 has a dense network of precise relations while in the 1853–1899 data only mentions parents and spouses; therefore, one of our future aims is to computationally extend the network to cover also the students of the 1853–1899 dataset.

The knowledge graph contains domain ontologies of 4800 place names and approximately 1000 links to our ontology of historical occupations AMMO [19]. The place ontology is the same as used in BiographySampo covering the most of Finnish towns and municipalities and also most frequently mentioned places abroad. The data will be further supplemented with ontologies of student nations and historical reference documents used as sources of information.

At this stage of work, the data has been manually validated, e.g., by making SPARQL queries or by converting the network to GraphML format [12] with the Python library NetworkX [13] and rendering it with software tools, such as Gephi [6]. Using the SPARQL queries we tested, e.g., that the years of people's birth, enrollment, and death with ages at each stage were all in a sensible range. This helped us to detect, e.g., false date or time span recognitions of our system.

A test version of the knowledge graph was published using the "7-star" Linked Data Finland model and platform (LDF.fi)[11]. LDF.fi extends Tim Berners-Lee's famous 5-star model[12] by two additional stars: the 6th start is given, if the dataset is published with the schemas it conforms to. The 7th start is given if an analysis of the quality of the data with respect to the schemas is provided, too [17]. An example of using the data service for visualizations is shown in Fig. 4, here the rich network of the family relationships of Karl Gustaf Ottelin (1792–1864).

[10] Women at the University of Helsinki (in Finnish) `http://www.helsinki.fi/yliopistonhistoria/yliopisto/nostot/naiset.htm`

[11] `http://ldf.fi`

[12] `https://5stardata.info/en/`

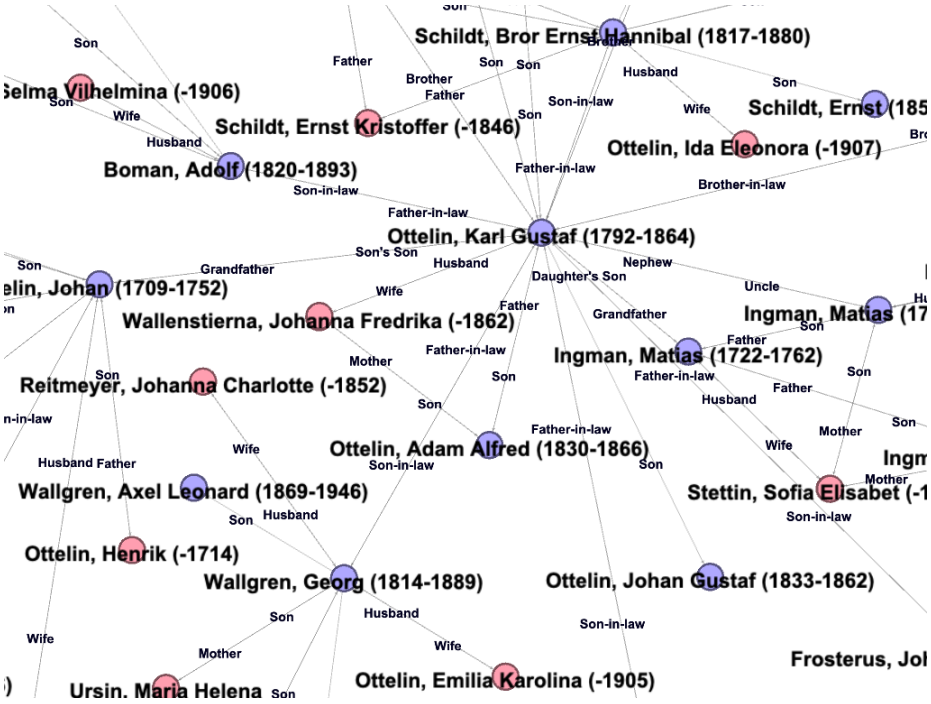


Fig. 4. Family relationships of the example person Karl Gustaf Ottelin

5 Related Work and Discussion

The Semantic Computing Research Group (SeCo) at Aalto University and University of Helsinki (HELDIG) has made earlier Linked Data publications collecting data about people in the history of Finland and beyond, including WarSampo on war history, BiographySampo, U.S. Congress Legislators [24], and Norssit High School Alumni registry [14]. The work of this paper is a continuation of these projects and further a part of a more comprehensive project of assembling an ontology of historical people in Finnish history, an important part of the emerging Linked Open Data Infrastructure for Digital Humanities in Finland initiative[13].

Representing and analyzing biographical data has grown into a new research and application field, reported, e.g., in the Biographical Data in Digital World workshops BD2015 [7], BD2017 [10], and BD2019. In [21], analytic visualizations were created based on U.S. Legislator registry data, and the Six Degrees of Francis Bacon system[14] [27,20] utilizes data of the Oxford Dictionary of National Biography. Extracting Linked Data from texts has been studied in several works, cf. e.g. [11]. In [9] language technology was applied for extracting entities and

[13] `https://seco.cs.aalto.fi/projects/lodi4dh/`
[14] `http://www.sixdegreesoffrancisbacon.com`

relations in RDF using Dutch biographies in the BiographyNet, as part of the larger NewsReader project [25].

Acknowledgements

Thanks to Yrjö Kotivuori and Veli-Matti Autio for their seminal work in creating the original Ylioppilasmatrikkeli databases used in our work, and for making the data openly available for our project to be published as Linked Open Data later on. Laura Sirola created a first RDF version of the datasets. CSC – IT Center for Science, Finland has provided computational server resources for our data service and applications.

References

1. American National Biography (2017), http://www.anb.org/aboutanb.html
2. Biography Portal of the Netherlands (2017), http://www.biografischportaal.nl/en
3. Dictionary of Swedish National Biography (2017), https://sok.riksarkivet.se/Sbl/Start.aspx?lang=en
4. Neue Deutsche Biographie (2017), http://www.ndb.badw-muenchen.de/ndb_aufgaben_e.htm
5. Nouvelle Biographie générale (2017), https://fr.wikipedia.org/wiki/Nouvelle_Biographie_generale
6. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. In: Third international AAAI conference on weblogs and social media (2009)
7. ter Braake, S., Anstke Fokkens, R.S., Declerck, T., Wandl-Vogt, E. (eds.): BD2015, Biographical Data in a Digital World 2015. CEUR Workshop Proceedings, Vol-1399 (2015), http://ceur-ws.org/Vol-1272/
8. Doerr, M.: The CIDOC CRM—an ontological approach to semantic interoperability of metadata. AI Magazine **24**(3), 75–92 (2003), https://doi.org/10.1609/aimag.v24i3.1720
9. Fokkens, A., ter Braake, S., Ockeloen, N., Vossen, P., Legêne, S., Schreiber, G., de Boer, V.: BiographyNet: Extracting Relations Between People and Events. In: Europa baut auf Biographien. pp. 193–224. New Academic Press, Wien (2017)
10. Fokkens, A., ter Braake, S., Sluijter, R., Arthur, P., Wandl-Vogt, E. (eds.): BD2017 Biographical Data in a Digital World 2015. CEUR Workshop Proceedings, Vol-1399 (2017), http://ceur-ws.org/Vol-2119/
11. Gangemi, A., Presutti, V., Recupero, D.R., Nuzzolese, A.G., Draicchio, F., Mongiovì, M.: Semantic web machine reading with FRED. Semantic Web Journal **8**, 873–893 (2017)
12. GraphML Team: The GraphML File Format, http://graphml.graphdrawing.org/ accessed: 30 September 2019
13. Hagberg, A., Swart, P., S Chult, D.: Exploring network structure, dynamics, and function using networkx. Tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States) (2008)
14. Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., Sirola, L.: Reassembling and enriching the life stories in printed biographical registers: Norssi high school alumni on the semantic web. In: Language, Technology and Knowledge. pp. 113–119. Springer–Verlag (2017)

15. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., Keravuori, K.: BiographySampo – publishing and enriching biographies on the semantic web for digital humanities research. In: Proceedings of the 16th Extended Semantic Web Conference (ESWC 2019). Springer–Verlag (2019)
16. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., 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) (September 2019)
17. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) The Semantic Web: ESWC 2014 Satellite Events. ESWC 2014. pp. 226–230. Springer-Verlag (May 2014), https://doi.org/10.1007/978-3-319-11955-7_24
18. Keith, T.: Changing conceptions of National Biography. Cambridge University Press (2004)
19. Koho, M., Gasbarra, L., Tuominen, J., Rantala, H., Jokipii, I., Hyvönen, E.: AMMO Ontology of Finnish Historical Occupations. In: Proceedings of the The First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH'19). vol. 2375, pp. 91–96. CEUR Workshop Proceedings (June 2019), http://ceur-ws.org/Vol-2375/, vol 2375
20. Langmead, A., Otis, J., Warren, C., Weingart, S., Zilinski, L.: Towards interoperable network ontologies for the digital humanities. Int. J. of Humanities and Arts Computing **10**(1), 22–35 (2016)
21. Larson, R.: Bringing lives to light: Biography in context (2010), Final Project Report, University of Berkeley, http://metadata.berkeley.edu/Biography_Final_Report.pdf
22. Leskinen, P., Hyvönen, E.: Extracting genealogical networks of linked data from biographical texts. In: Proceedings of the 16th Extended Semantic Web Conference (ESWC 2019), Posters & demonstrations (June 2019)
23. Leskinen, P., Tuominen, J., Heino, E., Hyvönen, E.: An ontology and data infrastructure for publishing and using biographical linked data. In: Proceedings of the Workshop on Humanities in the Semantic Web (WHiSe II). CEUR Workshop Proceedings (October 2017)
24. Miyakita, G., Leskinen, P., Hyvönen, E.: Using linked data for prosopographical research of historical persons: Case U.S. Congress Legislators. In: 7th International Conference, EuroMed 2018, Proc., Part II. pp. 150–162. Springer-Verlag (2018)
25. Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., Bogaard, T.: Building event-centric knowledge graphs from news. Web Semantics: Science, Services and Agents on the World Wide Web **37**, 132–151 (2016)
26. Tuominen, J., Hyvönen, E., Leskinen, P.: Bio CRM: A data model for representing biographical data for prosopographical research. In: Biographical Data in a Digital World (BD2017) (2017), https://doi.org/10.5281/zenodo.1040712
27. Warren, C., Shore, D., Otis, J., Wang, L., Finegold, M., Shalizi, C.: Six degrees of Francis Bacon: A statistical method for reconstructing large historical social networks. Digital Humanities Quarterly **10**(3) (2016)

Publication IX

Petri Leskinen and Eero Hyvönen. Reconciling and Using Historical Person Registers as Linked Open Data in the AcademySampo Portal and Data Service. *ISWC2021*, Andreas Hotho, Eva Blomqvist, Stefan Dietze, Achille Fokoue, Ying Ding, Payam Barnaghi, Armin Haller, Mauro Dragoni, Harith Alani (editors), pages 714—730, Springer, October 2021. online https://doi.org/10.1007/978-3-030-88361-4_42.

Reconciling and Using Historical Person Registers as Linked Open Data in the AcademySampo Portal and Data Service

Petri Leskinen[1][0000−0003−2327−6942] and Eero Hyvönen[1,2][0000−0003−1695−5840]

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

Abstract. This paper presents a method for extracting and reassembling a genealogical network automatically from a biographical register of historical people. The method is applied to a dataset of short textual biographs about all 28 000 Finnish and Swedish academic people educated in 1640–1899 in Finland. The aim is to connect and disambiguate the relatives mentioned in the biographies in order to build a continuous, genealogical network, which can be used in Digital Humanities for data and network analysis of historical academic people and their lives. An artificial neural network approach is presented for solving a supervised learning task to disambiguate relatives mentioned in the register descriptions using basic biographical information enhanced with an ontology of vocations and additional occasionally sparse genealogical information. Evaluation results of the record linkage are promising and provide novel insights into the problem of historical people register reconciliation. The outcome of the work has been used in practise as part of the in-use AcademySampo portal and linked open data service, a new member in the Sampo series of cultural heritage applications for Digital Humanities.

Keywords: Data Reconciling · Biographies · Linked Data · Digital Humanities.

1 Introduction

A key idea of Linked Data is to enrich datasets by integrating complementary local information sources in an interoperable way into a global knowledge graph. This involves harmonization of local data models used, as well as aligning the concepts and entities used in populating the local data models. The latter problem has been addressed traditionally in the field of *record linkage (RL)* [36,13,7], where the goal is to find matching data records between heterogeneous databases. For example, how to match person records in different registers, which may contain data about same persons, but where the data is represented using different metadata schemas and notational conventions? Using RL, richer global descriptions of persons can be created based on fusing local datasets. In addtion, RL facilitates data enrichment

by linking together local datasets that use different vocabularies and identifiers for representing same resources, such as persons.

This paper concerns the problem of entity reconciliation and RL of people in historical person registers. As a case study, academic people and their relatives extracted automatically from the textual biographical descriptions of the Royal Academy of Turku and University of Helsinki are considered. The primary data contains some 28 000 short biographical descriptions of people in 1640–1899, covering virtually all university students in Finland during this time period. This data contains not only the 1) the explicit set of students recorded but also 2) the implicit set of persons mentioned in the short biography record texts of (1), such as relatives and prominent historical persons. The task is to construct a knowledge graph of all persons referred to in the data (1)–(2) in order to study the characteristics of the underlying academic network.

As a solution approach, a probabilistic RL solution for linking person records is presented and tested with promising evaluation results. In our method, RL is based on the attributes of an actor, such as the name, life years, and vocations relating to her/his life. The key novel idea here is to enrich these attributes with genealogical information, i.e., information about the names and lifespans of actors' relatives. Integrating local person registers into a single global *knowledge graph (KG)* facilitates biographical and prosopographical research based on enriched data. For this purpose, the aligned enriched person data has been used as a basis for a new in-use semantic portal and data service, *AcademySampo – Finnish Academic People 1640–1899 on the Semantic Web*[3] [25].

This paper is structured as follows: We first present related works, the primary data of our study, and how it has been transformed into Linked Data. After this, the method of reconciling mentions of person in person registries is explained, and evaluation results in our case study are presented. In conclusion, contributions of the paper are discussed, and directions for further research are pointed out.

2 Related Work

The RL field is presented [13,36,3]. Several nation-wide projects are underway on integrating person registries. For example, the Norwegian Historical Population Register (HPR) is pursuing to cover the country's whole population in 1800–1964, based on combining church records and census data [32]. The Links project[4] in the Netherlands aims to reconstruct all nineteenth and early twentieth century families in the Netherlands based on civil certificates.

The problem of reconciling person records is evident in genealogical research. For example, in [26] Machine Learning has been applied to automatic construction of family trees from person records. Antolie et al. [2] present a case study

[3] The portal and its linked open data service, including a SPARQL endpoint, was released on February 5, 2021. More information about AcademySampo can be found on the project homepage: https://seco.cs.aalto.fi/projects/yo-matrikkelit/

[4] Cf. the project homepage https://iisg.amsterdam/en/hsn/projects/links and research papers at https://iisg.amsterdam/en/hsn/projects/links/links-publications.

of integrating Canadian World War I data from three sources: soldier records, casualty records, and census data. Here more traditional crafted RL processes were used, and using the data in research is demonstrated. Also Cunningham [8] concerns military person data. Here World War I military service records have been integrated with a census data, and the integrated data is used for data analysis. In Ivie et al. [19] the RL process is enhanced with the available genealogical data, e.g. information about spouses and children, to achieve a higher accuracy. Also Pixton et al. [27] utilize the genealogical information and apply a neural network for RL.

Representing and analyzing biographical data has grown into a new research and application field, reported, e.g., in the Biographical Data in Digital World workshops BD2015 [4], BD2017 [11], and BD2019. In [23], analytic visualizations were created based on U.S. Legislator registry data, and the Six Degrees of Francis Bacon system[5] [34,22] utilizes data of the Oxford Dictionary of National Biography. Extracting Linked Data from texts has been studied in several works, such as [12]. In [10], language technology was applied for extracting entities and relations in RDF using Dutch biographies in the BiographyNet, as part of the larger NewsReader project [29].

Our own earlier works related to the topic include reconciling biographees and their relatives in the BiographySampo semantic portal [15,24]. Here genealogical statistics e.g. average ages of becoming a parent or getting married were extracted from the source data, and person's life years are estimated according to that distribution. However, in this paper a neural network model is trained to learn similar rules from the data. References to World War II soldiers were reconciled for data linking in the in the WarSampo portal and knowledge graph [14,21].

3 Knowledge Graph of Historical Academic Persons

This section presents the data used in our study: the Finnish university student registries "Ylioppilasmatrikkeli" containing short biographical descriptions. A more detailed description about the data conversion is described in [25].

3.1 Primary Data Sources

The student registry datasets in our focus are based on original handwritten university enrollment documents. In an earlier project, the documents have been transliterated manually into textual form and extended with information from other sources about later life events of the biographees. It has been estimated that ten man years of manual work of archivists was needed to accomplish this.

Our work concerns two main parts of the student registry: the database covering the years 1640–1852[6] (*D1640*) available in Finnish and Swedish, and the registry of 1853–1899[7] (*D1853*) for the next years. The records contain

[5] http://www.sixdegreesoffrancisbacon.com
[6] https://ylioppilasmatrikkeli.helsinki.fi
[7] https://ylioppilasmatrikkeli.helsinki.fi/1853-1899

short biographical descriptions of 28 000 students of the University of Helsinki[8], originally the Royal Academy of Turku[9] in Finland. There are lots of mentions of relatives as well as of prominent related persons in the biographical descriptions. These student registries cover a significant part of the history of Finland and the Finnish university institution, since the University of Helsinki was the only university in the country during the time frame in focus. The data is widely used but genealogists and historians.

A key challenge in transforming this kind of data into Linked Data for data-analysis is how to reconcile mentions of people in the records and their biographical texts. For example, the data contains records of ten students with the same name *Johan Wegelius*. Furthermore, eight of them have a vocation related to clergy—more than half of the students who studied before the year 1780 worked as priests after their graduation.[10] In the textual descriptions of the students, there are 72 mentions of spouses or mothers with the name *Maria Johansdotter*. Furthernore, there are variations in how the names are written because the data has been collected from multiple sources by different archivists, when it was extended by additional information about the later lives of the students. For example, the name *Sofia Dorotea Cedercreutz* can also be written *Sophia Dorothea Cedercreutz*.

The data is divided into four parts: *D1640*: the students in 1640–1852; *R1640*: the relatives in *D1640*; *D1853*: the students of 1853–1899; *R1853*: the relatives of *R1853*. The aim is to link the people between these datasets. Therefore, the record linkage consists of the following partial tasks: 1) linkage from *R1640* to *D1640*, 2) linkage from *R1853* to *D1853*, 3) linkage from *R1853* to *D1640*, and 4) disambiguation of *R1640* and *R1853* data.

3.2 Extracting Information from Text

A comprehensive description about the data conversion as well as about the used data model is presented in an earlier article [25]. An extract of a registry entry for *Anders Israel Cajander*[11] is depicted in Fig. 1. The description starts with the date or year of enrollment, in this case *11.2.1830*. After that there is the full name and an unique database identifier followed by the place and time of birth (Leppävirralla 24.2.1811). Next there is a Finnish abbreviation *Vht* meaning parents; in the example case the father is Zachris Johan Cajander and the mother Gustava Karolina Neiglick. After that there are two lists of events, one related to studies and academic career, and other describing the later career of the biographee. At the end of the first paragraph, a person's death is marked with the symbol † and burial with ‡; the person in example died in Wyborg on December 18th, 1901 († Viipurissa 18.12.1901).

[8] https://en.wikipedia.org/wiki/University_of_Helsinki
[9] https://en.wikipedia.org/wiki/Royal_Academy_of_Turku
[10] This statistical result was obtained after we used the reconciled data in AcademySampo for data analysis.
[11] https://ylioppilasmatrikkeli.helsinki.fi/henkilo.php?id=14689

11.2.1830 **Anders Israel Cajander** 14689. * Leppävirralla 24.2.1811. Vht: Savon alisen kihlakunnan kruununvouti *Zachris Johan Cajander* (†1862) ja *Gustava Karolina Neiglick*. Kuopion triviaalikoulun oppilas 4.2.1822 – 22.6.1826 (betyg). Viipurin lukion oppilas 17.9.1827 – 1.7.1829. Ylioppilas Helsingissä 11.2.1830 (arvosana approbatur cum laude äänimäärällä 14). Viipurilaisen osakunnan jäsen 12.2.1830 *12/2 1830 \ Anders Israel Cajander \ 24/2 1811 \ KronoFogden Zachr. Joh. Cajander i Randasalmi \ Leppävirta \ [med betyg] fr. Gymn. i Wiborg \ Uttog betyg d. 12/10 1833 för att ingå vid Rättegångsverken.* Merkitty oikeustieteellisen tiedekunnan nimikirjaan 9.10.1832. Savokarjalaisen osakunnan perustajajäsen 1833 *Anders Israël Cajander.* Tuomarintutkinto 10.12.1833. Vaasan hovioikeuden auskultantti 24.12.1833. — Varatuomari 1837. Kihlakunnantuomarin arvonimi 1847. Äyräpään tuomiokunnan tuomari 1857, Jääsken tuomiokunnan 1870, Rannan tuomiokunnan 1877, ero 1891. Hovioikeudenasessorin arvonimi 1868. Laamannin arvonimi 1870. Valtiopäivämies 1872. †Viipurissa 18.12.1901.

Pso: 1841 *Fredrika Emelie Schildt* (†1892).
Veli: Räisälän kappalainen *Gustaf Adolf Cajander* 15376 (yo 1835, †1882).
Veli: kirjailija *Zakarias Cajander* 16147 (yo 1843, †1895).
Lanko: lääninmetsänhoitajan apulainen *Berndt Vilhelm Kristoffer Schildt* 14968 (yo 1832, †1892).

Fig. 1. Partial extract from a register entry text for *Anders Israel Cajander*

After the life time description, there are possible fields for relatives. In the example case, the spouse is mentioned first as Pso: 1841 Fredrika Emelie Schildt where *Pso* is a Finnish abbreviation for *puoliso* (spouse). There are three relatives who also have an entry in the register, i.e., two brothers (Veli: Gustaf Adolf Cajander and Veli: Zakarias Cajander) and a brother-in-law (Lanko: Berndt Vilhelm Kristoffer Schildt). The author of the *D1640* dataset Yrjö Kotivuori has manually added links from the description texts to the mentioned people also found in the register, like the three relatives in the example case. These links also contain linkage to the relatives in the *D1853* dataset.

3.3 Available Information

The previous person example was from the *D1640* data. However, the provided data in *D1853* differs in some aspects. For instance, *D1853* only mentions a person's parents and spouses, never children or any other relatives, and the people are not interlinked. Abbreviations are used generally for, e.g., vocations, which was taken into consideration in the data conversion by using specific lists of abbreviations.

Table 1 shows an analysis of the known positive sample pairs in the both datasets. Here column *source* refers to the relative, and *target* to the corresponding student entry. The rows show how many of the example pairs of particular data field are available, altogether the data contains 4285 training pairs. One can notice that for the six uppermost properties, e.g., preferred label, gender, death, vocation, child, and spouse are available for both the source and target records. On the other hand, the data fields indicating the place of death, year of birth, names of mother or father, as well as the alternative labels are usually not available. The column *common* indicates the number of cases where both the source and the target entries have the particular data field and *same* the number of entries where the source and the target values are equal.

This table clearly indicates which properties should be considered crucial in decision making. Notice that some attributes that are usually significant for a general case of RL, such as places of birth and death, are not chosen in this particular case study.

	Data 1640–1852				**Data 1853–1899**			
	source	target	common	same	source	target	common	same
preferable label	4285	4285	4285	3979	698	698	698	517
gender	4283	4284	4283	4283	698	698	698	688
year of death	4229	4208	4192	4141	135	352	134	130
vocation	4281	4270	4270	940	600	567	543	365
child	4285	4284	4284	3211	430	341	340	2
spouse	4285	4273	4273	-	698	687	687	2
place of death	2	3494	2	2	-	348	-	-
year of birth	-	2906	-	-	-	351	-	-
mother	-	3475	-	-	-	349	-	-
father	-	3478	-	-	-	348	-	-
alternative label	-	1761	-	-	30	165	29	22

Table 1. Available data fields in the training data

4 Method: Linking Person Records

This section describes the chosen formats for comparing two person registry entries. Generally, the input format for data comparison consists of numeric difference or similarity values between the data points of the two records, not the data of the records as it is. We first introduce the chosen input formats for data in different domains, e.g., for names and for vocations of the actors and the relatives. Finally, the architecture of the network model as well as the training setting are introduced.

4.1 Person Names

Person names in the datasets consists of a preferred and possibly alternative labels. Each label includes a family name and a sequence of given names. For the classifier input we considered four different variations of a label with a maximum of three first given names, only 0.4% of people entries have more than three given names. The classifier input is in a matrix format where the entry elements are statistical values calculated from the dataset. Each family and given name gets a *rarity* value so that first the frequency of the appearances for each name is

counted and the ranks are mapped into the numeric range $[0.0, 1.0]$: the most common names get a near-zero and the rarest values closer to 1.0 in order to distinguish the rare names.

Fig. 2 depicts an example of a name comparison matrix, in this case the family names of two person entries. The rows and columns mutually correspond to the data of two names that are compared. The uppermost row $(0.000, 0.808, 0.983, ...)$ consists of the rarity values for the first, and likewise the leftmost column $(0.000, 0.987, 0.991, 0.100, ...)$ for the second entry. The other values inside the matrix are Jaro-Winkler similarity values [35] between the name strings so that e.g,. perfectly matching names get the value 1.0.

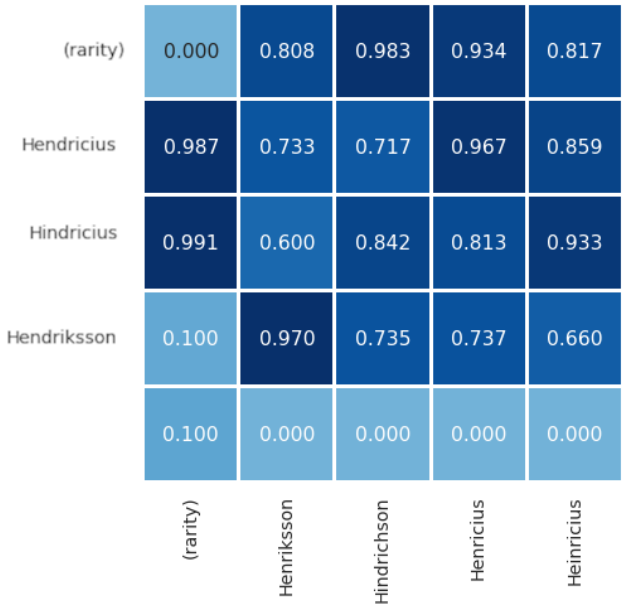


Fig. 2. Example of a matrix for comparing family names

4.2 Vocations

The vocations are the titles extracted from the source data. These titles often consists of a place name and a related profession, e.g., *Bishop of Turku* or *Bishop of Porvoo*. To enrich the data the vocations are linked to the hierarchical AMMO [20] ontology of historical occupations. Statistical values are used here like with the name entries. A *rarity* value is calculated for each title following the same principle as with the titles. In addition to that, a value of co-occurrences between two titles is calculated.

Fig. 3 depicts an example of a vocation comparison matrix. The value in the leftmost upper corner (0.455) is the Jaccard index [31] between the two sets of vocations. Similarly to the name matrix, the rows and columns correspond to the vocation in two dataset entries with the rarity values on the uppermost row $(0.909, 0.804, 0.249...)$ and leftmost columns. The rarity values are in a descending order so that the rarest vocations appear first on the lists. The other values filling the rest of the matrix are the co-occurrence values. In the data matrix, the co-occurrence value for a pair *(Law Reader, Mayor)* is 0.985, while the pair *(Court Attorney, Mayor)* has a value 0.250 indicating that this pair co-occurs in the data more frequently. The zero-valued elements on the right indicate that one of the title sets has less than the reserved seven data fields.

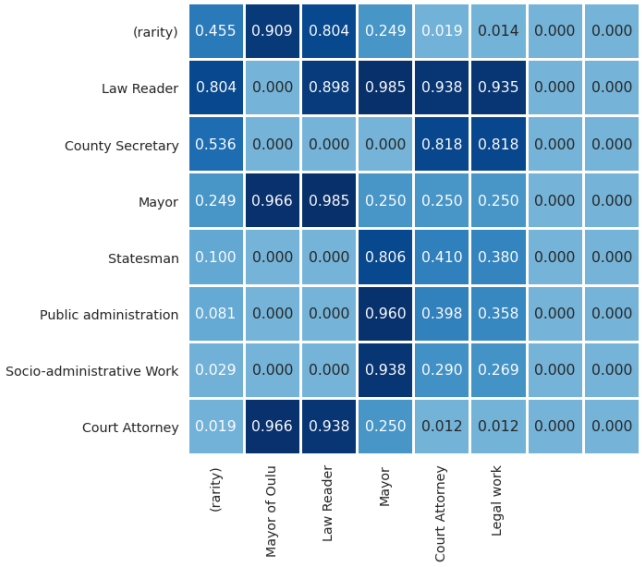


Fig. 3. Matrix for comparing the vocations

4.3 Years of birth and death, gender

The difference in actor's and relatives' birth and death years and their genders were also input to the network. The years use a precision of one year due to the format used in source data: the birth and death of the actor is usually known with a precision of a day, while in the case of relatives only the precision of a year is used. The actual difference is years is mapped into a near-zero range by using the arctan function. Gender was indexed using value -1.0 for female, 1.0 for male, and 0.0 for the rare cases where the gender was not known.

4.4 Relative information

The information of the relatives consists of details about the children and spouses of an actor, and basic information about her/his parents. The relative information uses the same matrix format as for the names, lifetime information, and vocations of the actor. It has reserved space for three children and three spouses, according to analyzing the data. In the data more than 99% have three or less spouses, and 95% three or less children.

4.5 Network model

The used network model is depicted in Fig. 4. It is a multi-input network based on the Keras functional API [6]. The network has eight inputs out of which six for the given and family names of the two actors, their spouses, and their children, one for the age comparison of actors and their relatives, and one for the actors' titles. The network acts as a probabilistic classifier and the output is $\bar{y} \approx [0.0, 1.0]$ for matching entries and $\bar{y} \approx [1.0, 0.0]$ for not matching pairs. For a binary decision these values are filtered by choosing the positive matches when the latter value exceeds a chosen threshold, e.g., $\lambda = 0.9$.

Some inputs are in a matrix format, which are first flattened[12], and after that run through a Dense[13] layer. Dropout layers with a ratio of 25% are used to prevent the overfit to the training data [30]. Different inputs of the same domain (e.g., names and years) are first concatenated[14] to one another. After a layer of Dense network the network concatenates into the final output.

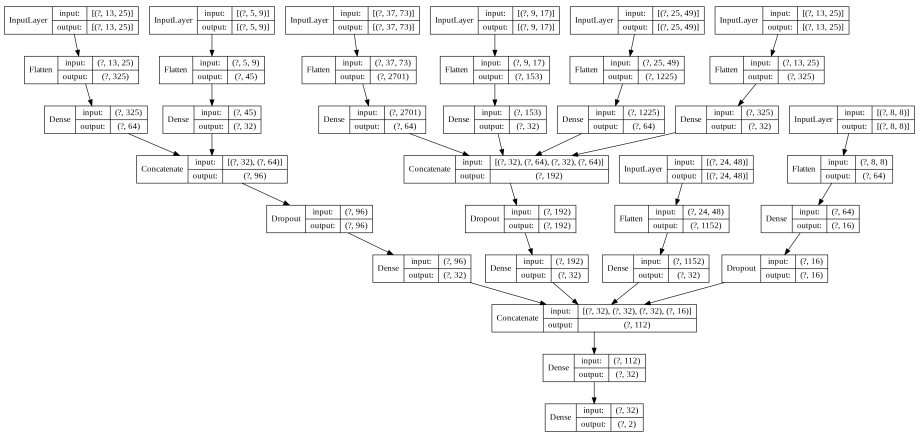


Fig. 4. Classifier model structure

[12] https://keras.io/api/layers/reshaping_layers/flatten/
[13] https://keras.io/api/layers/core_layers/dense/
[14] https://keras.io/api/layers/merging_layers/concatenate/

Training Data. The training data for the neural network could be input by either as an single data entry or in several smaller batches of data. We chose to feed the data in batches utilizing the Keras Data Generator Sequence [1] as described by A. Amidi and S. Amidi[15] due to the amount of data preprocessing from RDF format to numeric input.

Positive samples are created by reading the manually marked matches from the data. This linkage is many-to-one, so all the samples pointing to the same target can be chosen as training pairs pointing to each other. Finally, positive sample data is augmented with pairs where both the target and the source refer to the same resource.

The easiest way to gather the negative samples is to pick random pairs from the data. However, we chose to sample pairs that are likely to have some similar data values to improve the decision making. The dataset contains relations indicating, e.g., that two persons are siblings, cousins, or namesakes. Close relatives often have same similar characteristics, like family name or nearby years of birth.

Model Training. For the training the data was split into separate sets for training, testing, and validation of sizes 70%, 15%, and 15%, respectively. The classes in the training data are imbalanced, e.g., the number of negative samples ($N_n \approx 200000$) is significantly larger than the positive samples ($N_p \approx 13000$). Therefore the positive samples where defined to have a larger weight than the negative ones [33,5]. The training was performed in Google Colab, and the training with 100 epochs using a GPU took 4242.2 seconds. Validation accuracy of more than 99.6% was achieved during the training.

4.6 Evaluation

The results were analyzed closely by the Receiver Operating Characteristic (ROC) curve (Fig. 5) and by taking look at the details of False Positive and False Negative classifications. To deal with the data imbalance, a validation set with equal amount of positive and negative sample was used. The classifier input was divided by four different types: basic biographical information (B), genealogical information (G), name frequencies (N), and vocation frequencies (V). To analyze how much each data entry contributes to the prediction, evaluation was performed for four times using the entire data (B+G+N+V), biographical and genealogical data (B+G), biographical data with name and vocation frequencies (B+N+V), and the plain biographical data (B). The threshold value λ for optimal performance was chosen from the ROC curve coordinates by the point closest to the upper left corner [9]. For the entire data (B+G+N+V) the threshold value was $\lambda = 90.01\%$ and the resulting number of True Positives (TP) is 2035, True Negatives (TN) 2089, False Positives (FP) 0, and False Negatives (FN) 54 with measures precision of 100.00%, recall of 97.42%, F_1-score of 98.69%, and accuracy of 98.71%.

[15] https://stanford.edu/~shervine/blog/keras-how-to-generate-data-on-the-fly

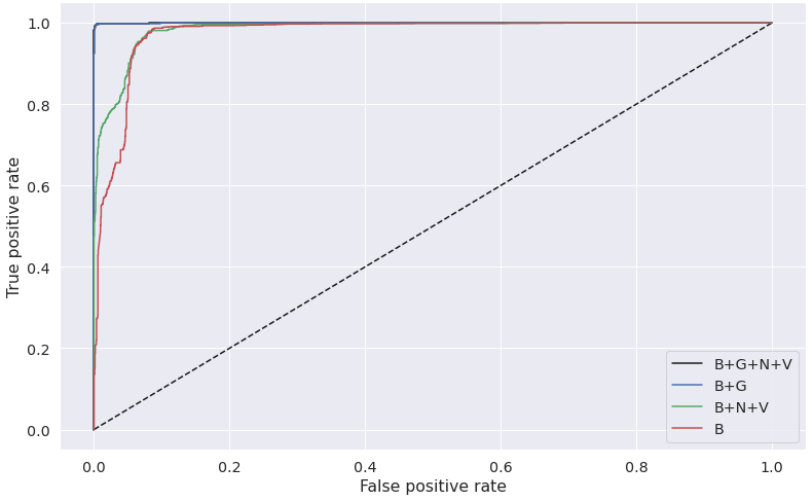


Fig. 5. ROC curve

In the ROC visualization, the curve with basic and genealogical (B+G) almost emerges with the curve for the entire data (B+G+N+V). Also Table 2 shows how close these results are to one another. Furthermore, the validation results without the genealogical information (B, B+N+V) show lower accuracy.

Data Subset	TP	FP	FN	TN	Precision	Recall	F_1-score	Accuracy	AUC	λ
B+G+N+V	2035	0	54	2089	100.00%	97.42%	98.69%	98.71%	99.98%	90.01%
B+G	2007	1	82	2088	99.95%	96.07%	97.97%	98.01%	99.97%	84.06%
B+N+V	2011	150	78	1939	93.06%	96.27%	94.64%	94.54%	98.47%	16.86%
B	587	12	1502	2077	98.00%	28.10%	43.68%	63.76%	97.48%	92.15%

Table 2. Validation results using different data subsets

Full disambiguation Record linkage with the real dataset was a many-to-one task, e.g. many records in the source set can be merged into one in the target data. When applying the model to the real dataset first blocking strategies [7] where applied to reduce the number of comparisons. For instance, candidate pairs of different gender or mismatching life years when known, could be omitted from candidate pairs. Likewise, candidates mentioned in a same register entry text e.g. siblings or different spouses could be omitted—same person is never mentioned twice in one text entry. Some preliminary disambiguation was performed already during the data conversion, e.g., aligning spouses of a person, if the names had a

high string similarity. The iterative process was run for several times because merging two person records furthermore can lead to finding more matches also among the relatives. To achieve a high precision and to minimize the number of false positive classification a high threshold values ($\lambda \geq 0.9$) were used.

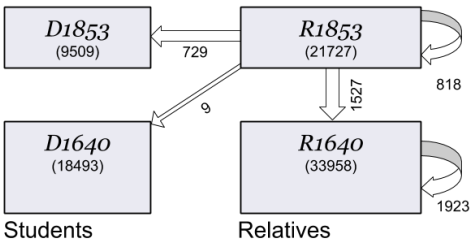


Fig. 6. Number of matches between the datasets

Fig. 6 depicts the number of records in each part of the dataset and the numbers of matches detected within them. The number of records before the RL are in parenthesis. For example, 729 of the records in *R1853* were merged into *D1853*, 1527 into *R1640*, and 9 into *D1640*. The latter number is relatively small because this matching was a part of the existing manual linkage by the dataset author, so these results are links missing from manual linkage or errors in our data conversion process. Inside the *R1853* dataset, 818 and in *R1640* 1923 entries were matched, respectively. Notice that we did not link the records from *R1640* to *D1640* because the existing manual linkage made by the dataset author.

5 Using AcademySampo

The people KG extracted from the primary data turned out be richly interlinked and forms the backbone of the AcademySampo portal and LOD service. Academic circles in history were smaller and people tended to marry within their own social class. For example, Fig. 7 depicts the extracted family relations of J. L Runeberg (1804–1877) (black large spot in the centre), the Finnish national poet, as visualized in one of the data-analytic views of the AcademySampo portal. Men in the figure are represented as blue and women as red spots. Most women in the data do not have a data entry of their own in the databases but are only mentioned in the biographies of the men because women were allowed to sign in universities only in the late 19th century. There are only 521 female academics out of 28 000 in the data.

The relations shown include both mentioned and inferred relations, such as brother in law, based on reasoning. Here is an example[16] of a SPARQL query

[16] https://api.triplydb.com/s/IE4w29n0T

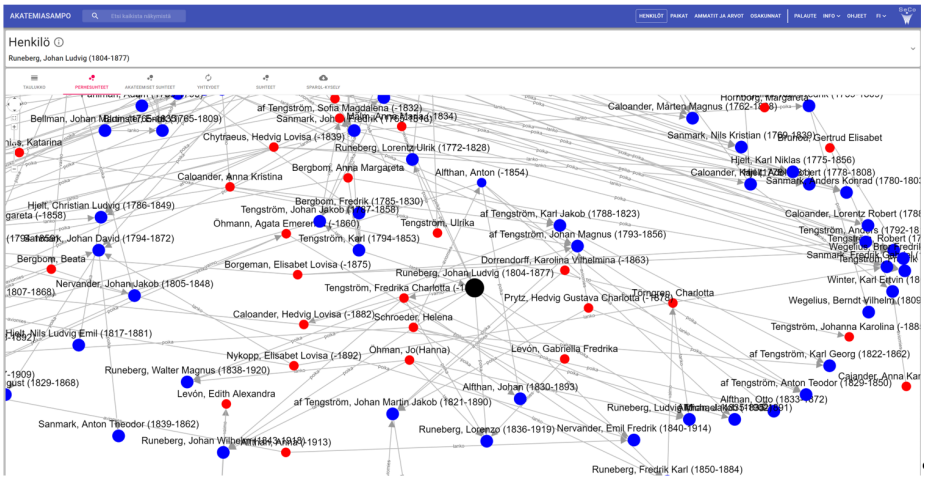


Fig. 7. Family relations of J. L Runeberg (1804–1877) visualized in AcademySampo

finds children of the same parent and concludes whether they are brothers or sisters based on the gender. Using AcademysSampo portal and the SPARQL endpoint for historical research is discussed in more detail in [17].

Deployment The AcademySampo KG was published on the Linked Data Finland platform[17] [16] powered by Fuseki SPARQL server[18] and Varnish Cache web application accelerator[19] for routing URIs, content negotiation, and caching. The portal user interface was implemented by the Sampo-UI framework [18]. AcademySampo system is based on Docker microservice architecture containers[20]. By using containers, the services can be migrated to another computing environment easily, and third parties can re-use and run the services on their own. The architecture also allows for horizontal scaling for high availability, by starting new container replicas on demand. The service has had 2300 users.

6 Discussion

The work described in this article shows that using genealogical information in RL is useful and can improve significantly the accuracy in person name reconciliation. This argument was tested and evaluated in detail in a case study using the AcademySampo datasets with promising results. We anticipate that similar results can be obtained in related use cases using other dataset. In the

[17] https://www.ldf.fi/dataset/yoma
[18] https://jena.apache.org/documentation/fuseki2/
[19] https://varnish-cache.org
[20] https://www.docker.com

AcademySampo project, the genealogical information has been used also when linking the records with Wikidata for semantic data enrichment.

When analysing the resulting matched pairs some weak cases needing separate handling where found. Historically, patronymic family names, e.g., *Johansdotter (Daughter of Johan)* have been common for women. However, the chosen Jaro-Winkler similarity may not be optimal to always disambiguate between cases like *Jöransdotter* and *Johansdotter*. Likewise, the classifier made some false results with the vocation of a farmer. Farmer was a common vocation in the 17th–19th century Finland, but yet rare in data records of academic people, for which reason we had put some excess weight on it in the classifying system.

This paper presented a method for reconciling person names mentioned in biographical texts of other people. The method was applied to creating a semantic KG of people that is used for studying and analyzing academic netwórks of people. For this purpose, the AcademySampo portal has been created, but also the underlying open linked data service can be used for custom-made data-analyses using, e.g., YASGUI[21] [28] and SPARQL or Python scripting in Google Colab[22] or Jupyter[23] notebooks, and for developing new applications [17].

Acknowledgements Thanks to Yrjö Kotivuori and Veli-Matti Autio for their seminal work in creating the original databases used in our work, and for making the data openly available. Discussions with Heikki Rantala, Esko Ikkala, Mikko Koho, and Jouni Tuominen are acknowledged. This work is part of the EU project InTaVia: In/Tangible European Heritage[24], and is related to the EU COST action Nexus Linguarum[25] on linguistic data science. CSC – IT Center for Science provided computational resources for the work.

References

1. Keras Documentation, Sequence. https://www.tensorflow.org/api_docs/python/tf/keras/utils/Sequence, accessed: 2020-12-10.
2. Antonie, L., Gadgil, H., Grewal, G., Inwood, K.: Historical Data Integration, a Study of WWI Canadian Soldiers. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). pp. 186–193. IEEE (2016).
3. Barlaug, N., Gulla, J.A.: Neural networks for entity matching. arXiv preprint arXiv:2010.11075 (2020).
4. ter Braake, S., Anstke Fokkens, R.S., Declerck, T., Wandl-Vogt, E. (eds.): BD2015, Biographical Data in a Digital World 2015. CEUR Workshop Proceedings, Vol-1399 (2015), http://ceur-ws.org/Vol-1272/.
5. Brownlee, J.: Machine Learning Mastery: How to Develop a Cost-Sensitive Neural Network for Imbalanced Classification. https://machinelearningmastery.com/cost-sensitive-neural-network-for-imbalanced-classification/, accessed: 2020-12-10.

[21] https://yasgui.triply.cc
[22] https://colab.research.google.com/notebooks/intro.ipynb
[23] https://jupyter.org
[24] https://intavia.eu/
[25] https://nexuslinguarum.eu/the-action

6. Chollet, F.: Keras, The Functional API. https://keras.io/guides/functional_api/, accessed: 2020-12-10.
7. Christen, P.: Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer Science & Business Media (2012).
8. Cunningham, A.: After "it's over over there": Using record linkage to enable the reconstruction of World War I veterans' demography from soldiers' experiences to civilian populations. Historical Methods: A Journal of Quantitative and Interdisciplinary History 51, 1–27 (2018).
9. Fawcett, T.: An introduction to ROC analysis. Pattern recognition letters 27(8), 861–874 (2006).
10. Fokkens, A., ter Braake, S., Ockeloen, N., Vossen, P., Legêne, S., Schreiber, G., de Boer, V.: BiographyNet: Extracting Relations Between People and Events. In: Europa baut auf Biographien. pp. 193–224. New Academic Press, Wien (2017).
11. Fokkens, A., ter Braake, S., Sluijter, R., Arthur, P., Wandl-Vogt, E. (eds.): BD2017 Biographical Data in a Digital World 2015. CEUR Workshop Proceedings, Vol-1399 (2017), http://ceur-ws.org/Vol-2119/.
12. Gangemi, A., Presutti, V., Recupero, D.R., Nuzzolese, A.G., Draicchio, F., Mongiovì, M.: Semantic web machine reading with FRED. Semantic Web 8, 873–893 (2017).
13. Gu, L., Baxter, R., Vickers, D., Rainsford, C.: Record linkage: Current practice and future directions. CSIRO Mathematical and Information Sciences (2003), cMIS Technical Report No. 03/83.
14. Heino, E., Tamper, M., Mäkelä, E., Leskinen, P., Ikkala, E., Tuominen, J., Koho, M., Hyvönen, E.: Named entity linking in a complex domain: Case second world war history. In: Proceedings, Language, Technology and Knowledge (LDK 2017). pp. 120–133. Springer–Verlag (June 2017), https://link.springer.com/chapter/10.1007/978-3-319-59888-8_10.
15. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., Keravuori, K.: BiographySampo – publishing and enriching biographies on the semantic web for digital humanities research. In: Proceedings of the 16th Extended Semantic Web Conference. Springer–Verlag (2019).
16. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) The Semantic Web: ESWC 2014 Satellite Events. ESWC 2014. pp. 226–230. Springer-Verlag (May 2014), https://doi.org/10.1007/978-3-319-11955-7_24.
17. Hyvönen, E., Leskinen, P., Rantala, H., Ikkala, E., Tuominen, J.: Akatemiasampo-portaali ja -datapalvelu henkilöiden ja henkilöryhmien historialliseen tutkimukseen (AcademySampo portal and data service for biographical and prosopographical research). Informaatiotutkimus (2021), accepted.
18. Ikkala, E., Hyvönen, E., Rantala, H., Koho, M.: Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. Semantic Web (2021), http://www.semantic-web-journal.net/, accepted.
19. Ivie, S., Pixton, B., Giraud-Carrier, C.: Metric-based data mining model for genealogical record linkage. In: 2007 IEEE International Conference on Information Reuse and Integration. pp. 538–543. IEEE (2007).
20. Koho, M., Gasbarra, L., Tuominen, J., Rantala, H., Jokipii, I., Hyvönen, E.: AMMO Ontology of Finnish Historical Occupations. In: Proceedings of the The First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH'19). vol. 2375, pp. 91–96. CEUR Workshop Proceedings (June 2019), http://ceur-ws.org/Vol-2375/, vol 2375.

21. Koho, M., Leskinen, P., Hyvönen, E.: Integrating historical person registers as linked open data in the warsampo knowledge graph. In: Semantic Systems. In the Era of Knowledge Graphs. SEMANTiCS 2020. LNCS, vol. 12378, pp. 118–126. Springer, Cham (2020), https://doi.org/10.1007/978-3-030-59833-4_8.
22. Langmead, A., Otis, J., Warren, C., Weingart, S., Zilinski, L.: Towards interoperable network ontologies for the digital humanities. Int. J. of Humanities and Arts Computing 10(1), 22–35 (2016).
23. Larson, R.: Bringing lives to light: Biography in context (2010), Final Project Report, University of Berkeley, http://metadata.berkeley.edu/Biography_Final_Report.pdf.
24. Leskinen, P., Hyvönen, E.: Extracting genealogical networks of linked data from biographical texts. In: Proceedings of the 16th Extended Semantic Web Conference (ESWC 2019). Springer–Verlag (2019).
25. Leskinen, P., Hyvönen, E.: Linked open data service about historical finnish academic people in 1640–1899. In: DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference. pp. 284–292. CEUR Workshop Proceedings, vol. 2612 (October 2020), http://ceur-ws.org/Vol-2612/short14.pdf.
26. Malmi, E., Gionis, A., Solin, A.: Computationally inferred genealogical networks uncover long-term trends in assortative mating. arXiv (2018), arXiv:1802.06055 [cs.SI].
27. Pixton, B., Giraud-Carrier, C.: Using Structured Neural Networks for Record Linkage. In: Proceedings of the Sixth Annual Workshop on Technology for Family History and Genealogical Research (2006).
28. Rietveld, L., Hoekstra, R.: The YASGUI family of SPARQL clients. Semantic Web – Interoperability, Usability, Applicability 8(3), 373–383 (2017). https://doi.org/10.3233/SW-150197.
29. Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., Bogaard, T.: Building event-centric knowledge graphs from news. Web Semantics: Science, Services and Agents on the World Wide Web 37, 132–151 (2016).
30. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks From Overfitting. The Journal of Machine Learning Research 15(1), 1929–1958 (2014).
31. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to data mining. 1st (2005).
32. Thorvaldsen, G., Andersen, T., Sommerseth, H.L.: Record linkage in the historical population register for Norway. In: Population reconstruction, pp. 155–171. Springer (2015).
33. Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., Kennedy, P.J.: Training deep neural networks on imbalanced data sets. In: 2016 international joint conference on neural networks (IJCNN). pp. 4368–4374. IEEE (2016).
34. Warren, C., Shore, D., Otis, J., Wang, L., Finegold, M., Shalizi, C.: Six degrees of Francis Bacon: A statistical method for reconstructing large historical social networks. Digital Humanities Quarterly 10(3) (2016).
35. Winkler, W.E.: String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage (1990).
36. Winkler, W.E.: Overview of Record Linkage and Current Research Directions. Tech. rep., U.S. Census Bureau (2006).

Publication X

Minna Tamper, Petri Leskinen, Jouni Tuominen, and Eero Hyvönen. Modeling and Publishing Finnish Person Names as a Linked Open Data Ontology. *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, June 2020, pages 3–14, ISSN 1613-0073, online `http://ceur-ws.org/Vol-2695/paper1.pdf` .

Modeling and Publishing Finnish Person Names as a Linked Open Data Ontology

Minna Tamper[1,2][0000−0003−1695−5840], Petri Leskinen[1][0000−0003−2327−6942], Jouni Tuominen[1,2][0000−0003−4789−5676], and Eero Hyvönen[1,2][0000−0003−1695−5840]

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

Abstract. This paper presents an ontology and a Linked Open Data service of tens of thousands of Finnish person names, extracted from contemporary and historical name registries. The repository, first of its kind available, is intended for Named Entity Recognition and Linking in automatic annotation and data anonymization tasks, as well as for enriching data in, e.g., genealogical research.

1 Introduction

Actor ontologies of people, groups, and organizations (e.g., Getty ULAN[3]), also called authority files [11] in Library Sciences, are a key ingredient needed in publishing and using Cultural Heritage (CH) Linked Data on the Semantic Web. For representing actor ontologies, there exists several vocabularies, such as FOAF[4], REL[5], BIO[6], and Schema.org [6]. Actor ontologies make a distinction between language-neutral concepts (resources identified by IRIs) and their literal names. In Resource Description Framework (RDF)[7]-based modeling in use on the Semantic Web, only resources can have properties while literal names are considered only atomic data that do not have properties, except a possible datatype and language tag attached. However, in many cases also literal words can have qualifiers and properties: names of things change in time and in context, e.g., female names due to marriage, or the language version form of the name in different countries and cultures (e.g., "Gabriela" vs. "Gabriele"). In linguistic Linked Data repositories [22], modeling phenomena related to the properties of words instead of real world things is actually the main reason for the research. For modeling phenomena like this, the SKOS recommendation has been extended

[3] http://www.getty.edu/research/tools/vocabularies/ulan/about.html
[4] http://xmlns.com/foaf/spec/
[5] http://vocab.org/relationship/
[6] http://vocab.org/bio/
[7] https://www.w3.org/RDF/

to SKOS-XL[8], allowing specifying properties for literal SKOS labels, and various linguistic ontology models such as Lemon[9] and OntoLex-Lemon[10] have been devised for representing linguistic Linked Data repositories.

A person name individualizes and identifies an individual. A person name ontology is a collection of contemporary and historical person names in a machine-understandable way. It is a knowledge graph describing names, their features, and usage in different datasets. In actor ontologies of people, names are often represented as literals. The features of the name are often ignored when describing people in actor ontologies although the name can carry information about its bearer such as socioeconomic status or gender.

This paper introduces a data model for representing person names as an ontology, based on tens of thousands of person names from contemporary Finnish name registries, including also historical names extracted from various CH linked data sources. The new Finnish Linked Open Data name ontology HENKO[11] has been used as a basis for named entity recognition (NER) and linking tasks [7] in automatic content annotation [29] and data anonymization services [25], as well as enriching linked data for applications, such as genealogical network analysis [16,20]. To foster the reuse of the data, this repository of Finnish person name data, first of its kind available, is published as a Linked Open Data service for application developers to use under the open CC BY 4.0 license.

2 Datasets

The data for the person name ontology HENKO was collected from multiple registries. It consists of given and family names and the number of users per name. The amount of users for the given names was calculated by gender. In addition, the given names data included the sum of users who have it as a first and as other given name. The collected datasets, the total number of names, and number of unique names in the data are shown in Table 1.

The first dataset in the table is from the Finnish Digital Agency[12] (FDA), a governmental agency that promotes digitalization of society, secures the availability of data, and provides services for the life events of its customers. The agency publishes Finnish name data as open data in the governmental publication portal avoindata.fi[13]. This dataset contains given names that are used by a minimum of five persons, and family names for the minimum of 20 persons. There are in total 23 018 family names, 9507 male given names, and 12 304 female given names (cf. Table 1). According to the product manager of FDA, the dataset contains only a fraction of Finnish person names. The full registry

[8] https://www.w3.org/TR/skos-reference/skos-xl.html
[9] https://lemon-model.net
[10] https://www.w3.org/2019/09/lexicog/
[11] The name comes from the Finnish name Henkilönimiontologia (Person name ontology); Henko is also a diminutive form of the name Henrik.
[12] https://dvv.fi/en/individuals
[13] https://www.avoindata.fi/data/en_GB/dataset/none

contains a total of 293 367 family names and 126 119 given names. According to FDA, the names used by less than the given amounts, are not publicly available because rare names can single out individual persons violating their privacy. Most of these unique names come from foreigners, and the rarer Finnish given names are often compound or coined names. FDA publishes the data twice a year; our the data has been collected starting from August 2018.

Dataset	**Family names**		**Given names**			**Total**
	unique	total	unique	female	male	
The Finnish Digital Agency	16 931	23 018	18 206	11 093	8299	42 410
BiographySampo	1205	5535	805	1705	1761	9001
Norssi High School Alumni	1002	4598	233	509	1039	6146
AcademySampo	6721	11 016	946	1389	1423	13 828

Table 1. Amount of names by dataset

In addition to using the FDA data, our ontology has names extracted from the datasets Norssi High School Alumni on the Semantic Web [9], BiographySampo [10], and AcademySampo [17][14]. AcademySampo contains names of university students from 1640 to 1899, and it contains plenty of historical, often Latin-based, names. BiographySampo data is based on 13 100 biographies of significant Finns throughout the history from the 3rd century to present time, and it has many Swedish names used by nobility and upper class because until 1809 Finland was an integral part of Sweden. The Norssi Alumni dataset records students in a Finnish school from 1867 to 1992 and the unique names in it are mostly rare Finnish names. Altogether these datasets provided 15 975 distinct family names, 2791 male and 2500 female names.

In order to have more features for the names in the ontology, the name datasets were processed and enriched using natural language processing (NLP) methods. Family names, for example, can contain nobiliary particles or suffixes. In Finnish family names [26] nobiliary particles are not used, but the names have suffixes that have indicated once if a person came from a place (e.g., suffixes *-la*, *-lä*), or person's socioeconomic status (e.g., scholars, soldiers, clergy with suffixes *-er*, *-ius*). To make this information explicit, the particles and suffixes were extracted from the names. For the particle extraction, the corpus of particles (in other languages) was compiled from the website of the Institute for the Languages of Finland (Kotus)[15] to identify names that contain particles. The extraction of suffixes was done using Lexical Analysis Service's (LAS)[16] [18,19] language recognition service, hyphenation service, and a manually compiled stopword list of words in Finnish and Swedish compound names

[14] `https://seco.cs.aalto.fi/projects/yo-matrikkelit/en/`
[15] `http://www.kielitoimistonohjepankki.fi/ohje/65`
[16] `http://demo.seco.tkk.fi/las/`

(e.g., fi. *Mansikkamaa* eng. *strawberry field*). The process first filters out names ending with a stopword, then detects the language, and lastly hyphenates the name. The last syllable is recorded as the suffix. Short names with only two syllables were ignored because they rarely end with a suffix.

In addition the NLP methods were used in identifying patronymics (e.g., *Jaakonpoika*, eng. *son of Jaakko*) and matronymics (e.g., *Liisantytär*, eng. *daughter of Liisa*). The matronymics and patronymics are identified in Finnish, Swedish, and Russian. In Finnish and Swedish they end with a word that indicates if its owner is a female (sv. *-dotter*, fi. *-tytär*) or a male (sv. *-son*, fi. *-poika*) whereas the Russian counterparts have a gendered suffix (e.g., *-ov*, *-ova*). The preceding part of the word is a person name typically in the genitive case and it can belong to an ancestor of the person. The ancestor's name wass extracted from the preceding part and baseformed with the LAS lemmatization tool. Afterwards, the ancestor's name is used to find names with the same string form. If the name exists, the application identifies the gender by using the existing data. The name instance is typed as matronymic or patronymic depending on the result.

Lastly, the names of HENKO data were linked to their counterparts in DBpedia and Wikidata to enrich the data with etymological information and relations to other names. The names were also linked to the bearers of the names in the source datasets (AcademySampo, BiographySampo, and Norssi Alumni). In addition, the family names can reference place and vocation names [26]. To identify names that refer to places and vocations, the names were linked to the YSO places ontology[17] (Finnish and Swedish place names) and to the Finnish historical occupations ontology AMMO [15]. This information is not only interesting topical information but can be used in tasks such as linking based NER to identify names that can be place or vocation names.

3 A Data Model for Person Names

The data model for person names in HENKO has been created based on the enriched name data. The data model is depicted in Fig. 1. The model has a class for the written representations of the name, the *WrittenNameForm*, that includes the string presentation of the name. Its instances are also instances of the CIDOC CRM's class *E41_Appellation* in order to enable the modeling of names and their alternative forms. This is needed, for example, if a name is translated from Russian to Finnish, as was the case with the Russian tsars *Alexandr I-III*, that were called in Finnish *Aleksanteri I-III*. The *WrittenNameForm* class connects to the *GivenName* and *FamilyName* classes via *isGivenName* and *isFamilyName* properties accordingly.

The *GivenName* and *FamilyName* classes are subclasses of the *Name* class. The *Name* class describes the basic features of the names, such as properties for linking both names to their equivalent representations in other ontologies, to people in other actor ontologies with the same names, in case of compound

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

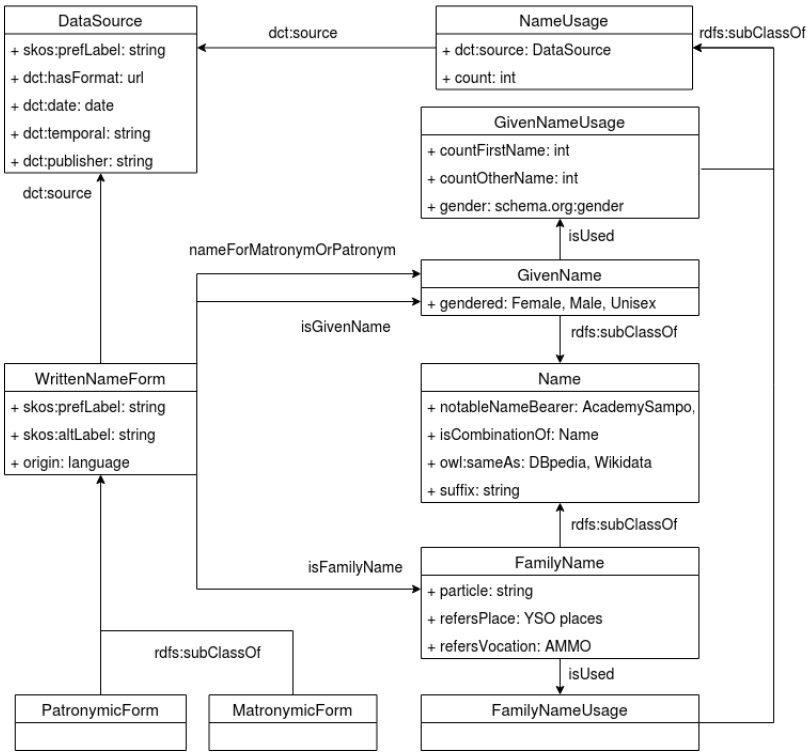


Fig. 1. The datamodel for Finnish person name ontology HENKO.

names by linking it to the parts (e.g., the name *Henna-Maria* can be linked to *Henna* and *Maria*), and to linguistic information, such as name suffixes. Like in the Wikidata [4] model, the *GivenName* class has information about the gender (male, female, unisex) that is inferred for each instance based on the name usage data. The *FamilyName* class instances contain information about the nobiliary particle, such as *von*, or *de la*. Initially, the OntoLex-Lemon [21] and MMoOn [14,13] ontologies were considered for modeling the particles and affixes, but the models did not fit the needs of HENKO because of being too complex or lacking in features to represent them. In addition, the references to places and vocations have been recorded using their own properties. Middle names[18] are not common in Finland and are ignored currently in the processing.

The *GivenName* and *FamilyName* classes are connected to the *GivenNameUsage* and the *FamilyNameUsage* classes through the *isUsed* property. These classes describe the calculated usage of the name. They are the subclasses of *NameUsage* class. The *NameUsage* class describes the general characteristics of its subclasses, such as count (how often a name is used) and source (data source for the information). The *GivenNameUsage* class also separates whether the name has been used as a first name or other name (second, third) in ad-

[18] `https://en.wikipedia.org/wiki/Middle_name`

dition to having the gender attribute. The *DataSource* class, that connects to the *NameUsage* superclass, describes the used sources in more detail. It includes attributes such as date (creation time of the data), URL (where the data was retrieved), temporal information about the dataset, its publisher, and name. The *DataSource* class is also connected directly to the *WrittenNameForm* class.

Finally, the *MatronymicForm* and *PatronymicForm* classes are subclasses of the class *WrittenNameForm*. If the instances of the *WrittenNameForm* class have been identified as patronymics or matronymics, the *WrittenNameForm* instances are complemented with information about the origin (Finnish, Swedish, Russian) and are linked to the given name of the ancestor (*GivenName* class instance) using using Wikidata's property "patronym or matronym for this name". The suffixes from the Russian origin names are recorded using the *Name* class property *suffix*.

4 Use Cases

This section presents the applications of HENKO in automatic annotation tasks. The applications are available as part of the SeCo Text Annotation Service[19].

Gender Identification Service The code behind the service[20] has been developed in the projects Norssi Alumni, BiographySampo, and AcademySampo to determine the gender by person's name. The service uses HENKO vocabularies of given names containing the frequencies of how often each name appears as a male or a female name.

The decision is based on the standard Bayesian approach described in equations 1 and 2. Equation 1 defines the probability $\rho(\gamma|n)$ that a person with a single given name n has gender $\gamma \in \{$"Female", "Male"$\}$. $D_F(name)$ and $D_M(name)$ are the frequencies of the $name$ in the vocabularies of female D_F and male D_M names. The smoothing variable α prevents the probabilities from getting near-zero values in ambiguous cases. In this way, e.g., names with only a few samples do not affect the final result too much. Likewise, if a name does not appear in either vocabulary, the estimate reduces to 50%—a natural choice for a prior probability when estimating an unknown gender. Equation 2 defines the probability that a given sequence of names $N = (name_1, name_2, \ldots)$ relates to gender γ. To simplify the calculations, the correlation between the names in the sequence was theorized to be statistically independent, e.g., having $name_1$ would not correlate with having $name_2$. Besides, the used vocabularies do not include information about the co-occurrences of given names. Therefore the probability of a sequence could be calculated as a product of the probabilities for each name.

$$\rho(\gamma|name) = \frac{\rho(name|\gamma) \cdot \rho(\gamma)}{\rho(name)} \approx \frac{D_\gamma(name) + \alpha}{D_F(name) + D_M(name) + 2\alpha} \quad (1)$$

[19] `https://nlp.ldf.fi`

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

$$\rho(\gamma|N = (name_1, name_2, \ldots)) = \frac{\prod\limits_{n \in N} \rho(\gamma|n)}{\prod\limits_{n \in N} \rho(\text{"Female"}|n) + \prod\limits_{n \in N} \rho(\text{"Male"}|n)} \quad (2)$$

For the final decision making, a threshold value τ (e.g., $\tau = 0.75$) is used. For example, if $\rho(\text{"Female"}|N) > \tau$, then the person is classified as a female, or as a male in case $\rho(\text{"Male"}|N) > \tau$. Moreover, no inference is made in the range $\rho \in [1.0-\tau, \tau]$ where the gender remains undefined. For example, when analyzing a unisex name like *Dominique*, the result remains undefined, but adding another name *Gaston*, the application interprets the sequence *Dominique Gaston* as a male name, or as a female in the case *Gabrielle Dominique*.

Person Name Finder Service The Person Name Finder is an API service for identifying references to people and collecting context around them from texts. It utilizes the HENKO ontology to identify person names from texts as a NER task. The Person Name Finder uses the linkage of the family names to places and vocations to differentiate between them and person names. In case the application finds from a text a reference to a single family name and there are no full names with the same family name in the text, it checks if the name is linked to either a place or vocation. If the family name has been linked to a place name, the application returns the place reference to indicate that the name can also be a place. The same procedure is applied to vocations; if a sentence starts with a name that is linked to a vocation written with a capital letter in a beginning of a sentence, the application returns the vocation link. Otherwise, the application returns only person names with links to the person name ontology. In addition, the service can identify information around the name such as times of birth and death, and the gender by utilizing the Gender Identification Service.

The service identifies person names and returns the result set in JSON format. It has been designed to aid in the extraction of personal information from registry entries and natural language texts. The result set contains full names and offers information related to the name such as location in text, links to HENKO, and optionally contextual information, such as gender, dates within brackets, etc. The API and its description[21] are available at the SeCo Text Annotation Service. Currently, the application is being developed and used as a part of named entity recognition and linking to identify person names from the legal and biographical texts. It has been able to identify most names and even some older names, and to enrich them with information such as years within brackets, and gender.

5 Evaluation

This section evaluates the enriching methods for the initial data in Section 2 and the Gender Identification Service from Section 4.

[21] `http://nlp.ldf.fi/api-documentation/#api-NameFinder`

The use of NLP methods for data enrichment provided satisfactory results. The identification of matronymics and patronymics was calculated for 1000 random samples. The F1-score for identification of matronymics was 87.27% and for patronymics 94.42%. Most frequently encountered issue with identification was the lack of Swedish or Russian given names from which the form is derived from. The extraction of suffixes and particles worked well. The F1-score for a sample of 1000 names was 92.78% for suffixes and 100% for particles. The suffix extraction failed for rarer non-Finnish names because they could not be hyphenated correctly due to language identification or lack of hyphenation support.

The linking of names succeeded with varying results. Roughly 23 600 names are linked to Wikidata, and 2500 to DBpedia. The rest of the names could not be linked because either the database did not include the name or there were errors in the data. Often older or less popular names could not be found in either target ontology. Also, some Asian names were linked to several entities in Wikidata with the same label, e.g. *Jin* was linked to two Chinese and one Korean name. The linking of names to topics matched to 785 places and 30 vocations. The success of the linking depended on the quality and coverage of the target ontology. Names from pre-Christian era could not be linked to places or vocations because the target ontologies do not contain a historical vocabulary for the entities.

The Gender Identification Service was evaluated using the names of the relatives extracted from BiographySampo data. It recognized 97.70% of the unique names leaving out only very rare or foreign names. In the test set, all recognized genders were inferred correctly [16]. In addition to using given names, the gender can be concluded e.g. by occupation, by known family relations, or by external contextual information. For example, in the case of AcademySampo all students starting earlier than in 1870 are male [17] since female students were not allowed.

6 Data Service

The person name ontology is published as Linked Open Data on the Linked Data Finland (LDF.fi) platform [8], adhering to the FAIR principles[22]. The platform provides a public SPARQL endpoint[23], IRI dereferencing capabilities, including a generic RDF browsing user interface, and a dataset homepage[24] with general documentation based on the SPARQL Service Description[25], containing a Vocabulary of Interlinked Datasets (VoID) description[26] of the dataset. For human-readable data model documentation[27], we use LODE [27]: when dereferencing IRIs of the name ontology's schema, the user is redirected to a page listing the classes and properties used. The ontology is also published in the ONKI Light

[22] `https://www.go-fair.org/fair-principles/`
[23] `http://ldf.fi/henko/sparql`
[24] `http://ldf.fi/dataset/henko`
[25] `https://www.w3.org/TR/sparql11-service-description/`
[26] `https://www.w3.org/TR/void/`
[27] `http://ldf.fi/schema/henko/`

service[28], where it is searchable and browsable using SKOSMOS[29], a web-based SKOS browser. The data is served on the Apache Jena Fuseki triplestore. The Fuseki runtime and the person name ontology data are built into a Docker image[30] which can be easily rebuilt when there is a need to publish a new version of the data, by simply updating the data in a Git repository.

7 Conclusions

This paper presents the person name ontology HENKO that consists of Finnish person names from the 3rd century to present time. Unlike actor ontologies and vocabularies such as ULAN and BIO, HENKO concentrates on describing person names and their features. The ontology is published as linked open data that connects to AcademySampo, BiographySampo, Norssi Alumni datasets and semantic portals, Wikidata, DBpedia, YSO places, and AMMO ontologies. Its unique data model was influenced by largely used ontologies and vocabularies such as Wikidata, Schema.org, and DBpedia. Out of these ontologies, Wikidata has the most extensive model thus far for names; it divides names by gender, includes etymological information, and has pronunciation instructions. In addition, the Wikidata ontology differentiates patronymic and matronymic names. In contrast, HENKO consists of a large set of Finnish names of which nearly 45% could be linked to Wikidata. In addition, the HENKO has more information about the names such as their usage statistics, linguistic information (suffixes, particles), and provenance information. HENKO model can be used as is for simple patterns consisting of given and family names. In addition, by adding the modelling for middle names, it can be used for wider range of naming conventions. Hence, the ontology is a novel resource for different applications. It can also be used as training material for deep learning based NLP applications alike.

The accuracy of extracting particles and suffixes was satisfactory. The minor issues of suffix extraction could be solved by identifying and splitting family names that are compound words with tools such as the Turku dependency parser [12] or LAS's morphological analyzer. In addition to family names, also given names can contain suffixes that have so far been ignored. They can, e.g., indicate the bearer's gender, like in the female Wilhelmiina based on the male name Wilhelm. The identification and extraction of suffixes enables data analysis for the names. For example, in the history of Finnish last names [26], there have been periods when it has been popular to change Swedish or Russian names to Finnish names with suffixes such as *-la* or *-nen*. When analyzing the AcademySampo data, we found out that family names with suffix *-nen* start to appear only after 1830. To analyze the temporal characters of family names with other suffices remain as future work. Given names [28] have also been modified but by the clergy keeping the parish registries according to the guidelines of different central governments; for example the name Gregorius has been changed to the

[28] `http://light.onki.fi/henko/en/`
[29] `http://skosmos.org`
[30] `https://hub.docker.com/r/secoresearch/fuseki/`

Finnish name Reijo[31]. One future research direction for enriching the data could be to represent there changes of names based on genealogical data and track the changes and suffixes in different linked source datasets. This would also aid in named entity linking (NEL), as the name changes in historical documents could be understood and references to people could be disambiguated better if indicated that the person used different changed names. Modeling of the changes of names has been researched earlier, e.g., in the context of biological taxa [30,3].

The linking of family names to places and vocations enriched the ontology and added context to names. The Person Name Finder utilizes the added context to identify possibly ambiguous nouns when it is used to identify names from text. Unlike typical NEL tools [23,24,5] that concentrate on simply linking entities to knowledge bases, the application can be utilized to extract names from texts and enrich them with contextual information. The Person Name Finder application is still under work, and will be further developed to ease linking to related actor ontologies. In addition to topical linking, in the future, place name linking can be used similarly to, e.g., Tuomas Salste's work[32] by locating the origin of names and visualizing them on a map to aid in genealogical research. By using the extracted suffixes, the linking of names to places could be improved and expanded to names that refer to places but contain a suffix that prevents linking (e.g., Savola referes to Savo without the -la suffix).

The usage statistics of the names enables the usage of the ontology in the Gender Identification Service. Although the functionality of the service is straightforward and based on relative trivial statistics, e.g., it does not consider the co-occurrence of the names and it does not return an estimate for names missing in the ontology, the results have been feasible in our use cases. Related to our service, there are commercial projects such as genderize[33] and gender-api[34] that also use name vocabularies for decision making. Attempt to infer the gender by the ending of the name [1] is problematic with Finnish names where, e.g., *Jari* and *Kari* are male names but *Sari* and *Mari* female ones. A blog post [2] by Ellis Brown introduces a project where the gender is inferred from character sequences in names using a recurrent neural network. Due to the feasible results for our use cases, we have not implemented similar algorithms for inferring the gender for names missing from our vocabulary.

Acknowledgments This work is part of the Anoppi project[35] funded by the Ministry of Justice in Finland. Thanks 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 University of Helsinki and Esko Kirjalainen from The Finnish Digital Agency for insightful discussions. CSC – IT Center for Science, Finland, provided us with computational resources.

[31] `https://www.genealogia.fi/nimet/nimi15s.htm`
[32] `https://www.tuomas.salste.net/suku/nimi/`
[33] `https://genderize.io`
[34] `https://gender-api.com`
[35] `https://seco.cs.aalto.fi/projects/anoppi/en/`

References

1. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit, chap. 6.1.1. O'Reilly Media, Inc. (2009)
2. Brown, E.: Gender Inference from Character Sequences in Multinational First Names. https://towardsdatascience.com/name2gender-introduction-626d89378fb0, accessed: 2020 Mar 3
3. Chawuthai, R., Takeda, H., Wuwongse, V., Jinbo, U.: Presenting and Preserving the Change in Taxonomic Knowledge for Linked Data. Semantic Web **7**(6), 589–616 (2016)
4. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D.: Introducing Wikidata to the linked data web. In: International Semantic Web Conference. pp. 50–65. Springer (2014)
5. Francis-Landau, M., Durrett, G., Klein, D.: Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks. arXiv preprint arXiv:1604.00734 (2016)
6. Guha, R.V., Brickley, D., Macbeth, S.: Schema.org: Evolution of Structured Data on the Web. Communications of the ACM **59**(2), 44–51 (2016)
7. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R.: Evaluating entity linking with Wikipedia. Artificial Intelligence **194**, 130–150 (Jan 2013). https://doi.org/10.1016/j.artint.2012.04.005, http://dx.doi.org/10.1016/j.artint.2012.04.005
8. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets. In: The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers. pp. 226–230. Springer-Verlag (May 2014)
9. Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., Sirola, L.: Reassembling and Enriching the Life Stories in Printed Biographical Registers: Norssi High School Alumni on the Semantic Web. In: Proceedings, Language, Technology and Knowledge (LDK 2017). pp. 113–119. Springer-Verlag (June 2017)
10. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., 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) (September 2019)
11. Joudrey, D., Taylor, A., Miller, D.: Introduction to Cataloging and Classification. Libraries Unlimited, 11 edn. (2015)
12. Kanerva, J., Ginter, F., Miekka, N., Leino, A., 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. Association for Computational Linguistics (2018)
13. Klimek, B.: Proposing an OntoLex-MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models. In: LDK Workshops. pp. 68–73 (2017)
14. Klimek, B., Arndt, N., Krause, S., Arndt, T.: Creating Linked Data Morphological Language Resources with MMoOn – The Hebrew Morpheme Inventory. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 892–899 (2016)
15. Koho, M., Gasbarra, L., Tuominen, J., Rantala, H., Jokipii, I., Hyvönen, E.: AMMO Ontology of Finnish Historical Occupations. In: Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH'19). vol. 2375, pp. 91–96. CEUR Workshop Proceedings (June 2019), vol 2375

16. Leskinen, P., Hyvönen, E.: Extracting Genealogical Networks of Linked Data from Biographical Texts. In: Proceedings of the 16th Extended Semantic Web Conference (ESWC 2019), Portoroz, Posters & Demonstrations (June 2019)
17. Leskinen, P., Hyvönen, E.: Linked Open Data Service about Historical Finnish Academic People in 1640–1899. In: Proceedings of Digital Humanities in Nordic Countries (DHN 2020), Riga. CEUR Workshop Proceedings (March 2020)
18. Mäkelä, E.: Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text. In: European Semantic Web Conference. pp. 424–428. Springer (2014)
19. Mäkelä, E.: LAS: an integrated language analysis tool for multiple languages. The Journal of Open Source Software **1**(6) (October 2016)
20. Malmi, E., Rasa, M., Gionis, A.: AncestryAI: A Tool for Exploring Computationally Inferred Family Trees. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 257–261. International World Wide Web Conferences Steering Committee (2017)
21. McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P., Cimiano, P.: The OntoLex-Lemon model: development and applications. In: Proceedings of eLex 2017 conference. pp. 19–21 (2017)
22. McCrae, J.P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., De Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., et al.: 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). pp. 2435–2441 (2016)
23. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th international conference on semantic systems. pp. 1–8. ACM (2011)
24. Nguyen, D.B., Hoffart, J., Theobald, M., Weikum, G.: AIDA-light: High-throughput named-entity disambiguation. In: Proceedings of the Workshop on Linked Data on the Web (LDOW 2014), co-located with the 23rd International World Wide Web Conference (WWW 2014). vol. 1184. CEUR Workshop Proceedings (April 2014)
25. Oksanen, A., Tamper, M., Tuominen, J., Hietanen, A., Hyvönen, E.: Anoppi: A pseudonymization service for Finnish court documents. In: Araszkiewicz, M., R.D.V. (ed.) Legal Knowledge and Information Systems. JURIX 2019: The Thirty-second Annual Conference. pp. 251–254. IOS Press (December 2019)
26. Paikkala, S.: Sukunimet sukututkimuksessa. Sukutieto: Sukutietotekniikka ry:n jäsenlehti **14**(4) (1997)
27. Peroni, S., Shotton, D., Vitali, F.: Tools for the Automatic Generation of Ontology Documentation: A Task-Based Evaluation. International journal on Semantic Web and information systems **9**(1), 21–44 (2013)
28. Rajasuu, R.: Kuopiossa, Oulussa ja Turussa vuosina 1725–1744 ja 1825–1844 syntyneiden kastenimet. Ph.D. thesis, University of Eastern Finland (2013)
29. Tamper, M., Hyvönen, E., Leskinen, P.: Visualizing and Analyzing Networks of Named Entities in Biographical Dictionaries for Digital Humanities Research. In: Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICling 2019). Springer-Verlag (April 2019)
30. Tuominen, J., Laurenne, N., Hyvönen, E.: Biological Names and Taxonomies on the Semantic Web – Managing the Change in Scientific Conception. In: Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011). pp. 255–269. Springer-Verlag (June 2011)

Publication XI

Petri Leskinen, Javier Ureña-Carrion, Jouni Tuominen, Mikko Kivelä, and Eero Hyvönen. Knowledge Graphs and Data Services for Studying Historical Epistolary Data in Network Science on the Semantic Web. *Submitted for review*, Semantic Web Journal .

Semantic Web 0 (2023) 1–0
IOS Press

Knowledge Graphs and Data Services for Studying Historical Epistolary Data in Network Science on the Semantic Web

Petri Leskinen [a,b,*], Javier Ureña-Carrion [c], Jouni Tuominen [a,b,d], Mikko Kivelä [c] and Eero Hyvönen [a,b]
[a] *Semantic Computing Research Group (SeCo), Aalto University, Finland*
E-mails: petri.leskinen@aalto.fi, jouni.tuominen@aalto.fi, eero.hyvonen@aalto.fi
[b] *HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland*
[c] *Complex Systems Group, Aalto University, Finland*
E-mails: javier.urenacarrion@aalto.fi, mikko.kivela@aalto.fi
[d] *HSSH – Helsinki Institute for Social Sciences and Humanities, University of Helsinki, Finland*

Abstract. Communication data between people is a rich source for insights into societies and organizations in areas ranging from research on history to investigations on fraudulent behavior. These data are typically heterogeneous datasets where communication networks between people and the times and geographical locations they take place are important aspects. We argue that these features make the area of temporal communications a promising application case for Linked Data (LD) -based methods combined with temporal network analyses. The key result of this paper is to present a framework, tools and systems, for creating, publishing, and analyzing historical LD from a network science perspective. The focus is on network analysis of epistolary network data (metadata about letters), based on recent advances in analysis of temporal communication networks and the behavioral patterns commonly found in them. To test, evaluate, and demonstrate the usability of the framework, it has been applied to (1) the Dutch CKCC corpus (of ca. 20 000 letters), (2) the pan-European correspSearch corpus (of ca. 135 000 letters), (3) to the Early Modern Letters online data (of ca. 160 000 letters), and (4) to the aggregated Finnish CoCo collection of more than 300 000 letters from 1809–1917.

Keywords: Semantic Web, Linked Open Data, Digital Humanities, Network Science, Early Modern

1. Introduction

Since the revolution in network science around 20 years ago [34, 40, 41], this field of research has been extremely successful explaining various phenomena and fundamental concepts in a wide array of systems from societies to brain and cellular biology. The tools and ideas developed for network analysis allow for different levels of granularity ranging from the whole network to diagnostics computed for individual nodes in the network, such as centrality measures, node roles, and local clustering coefficients. However, these tools are often mainly used by the network scientists as they are difficult to use for the domain experts: accessing them requires programming skills or at least specialised software that relates the often heterogeneous network data and metadata to the questions that are important for the domain experts. On the other hand, there is a need to make the rich datasets created by historians in Digital Humanities (DH) and the Linked Data community available for the network scientists.

*Corresponding author. E-mail: petri.leskinen@aalto.fi.

This paper builds on the idea that Semantic Web technologies[1] [12] and Linked Data [9, 16] can be a solution to these problems. The graph-based RDF data model underlying the Semantic Web is a perfect match for representing network data, and Linked Data publishing [9] can be used for making the data available for researchers in humanities with some skills on using SPARQL[2] queries or on programming with SPARQL endpoints. Furthermore, ready-to-use portal solutions for data analysis can be implemented for DH based on such data services [19]. The idea is that by combining the flexibility of publishing and using LD with the tools of network science can help domain experts to tackle massive network data in fruitful manner with little or no expertise in programming. Furthermore, the LD created can be served back to the research community for further research and application development in a disciplined and well-defined way by using the Semantic Web methodology [13] with practical LD publishing principles including SPARQL endpoints.

Table 1

Datasets analyzed and discussed in this paper

Dataset	Content
1. CKCC	Epistolary data of the CKCC corpus of the Huygens Institute in the Netherlands, an aggregated collection of ca 20 000 Dutch correspondences [10, 27] related to the Republic of Letters [15, 27]
2. correspSearch	Epistolary data 1510–1991 of 135 000 letters aggregated by the correspSearch project at the Berlin Brandenburg Academy of Sciences and Humanities [4]
3. EMLO	Early Modern Letters Online (EMLO) dataset was transformed from our Linked Data version [36] of the aggregated Early Modern Letters Online Database EMLO at Oxford University. The dataset includes metadata about ca. 160 000 letters sent in 1500–1800.
4. CoCo	Epistolary data about letters sent in the Grand Duchy of Finland 1809–1917, harvested by the project *Constellations of Correspondence (CoCo)*[3] [35].

To test and demonstrate this approach in practise, this paper focuses on communication networks that are represented as temporal networks, a rapidly developing subfield of network science [14, 34]. Thedatasets of historical epistolary data listed in Table 1 are used for case study examples. Temporal networks are a specific type of networks that carry information on the activation times of the links in addition to the topological structure of the networks. In communication networks this means that we do not only consider who has been in contact with whom, but also the exact time instances at which the communication has taken place. This not only adds complications related to how the various methods and measures are generalised for temporal networks, but also creates possibilities of new types of network analysis. For example, in communication networks it has been found that the individuals are in contact in a bursty manner [6, 22] and they distribute their communication efforts via patterns, known as social signatures, that are specific to each individual [11, 33]. These phenomena are understood in terms of statistical laws found in anonymized data, but much less attention has been given on how such features translate to interpretations of individual relationships or people. Here we introduce a method for giving access to these state-of-the-art network analysis methods to domain experts, who work through the massive databases of communications using theoretically grounded analysis tools.

The paper extends our earlier papers related to publishing and analyzing historical epistolary data and LetterSampo [17, 37] by the network science perspective outlined above, and by presenting tools and systems for network analysis. The linked open data resources regarding datasets 1 and 2 of Table 1 are available online both as data dumps in Zenodo.org and in a SPARQL endpoint, as described in more detail in [17]. The idea of the LetterSampo data service and the portal, and how the tools are used, are illustrated in this online video[4].

It should be noted that this paper focuses on presenting a technical framework and approach for applying network analysis and LD technology to publishing and using historical epistolary data in research, not on particular domain specific analyses of the datasets from a humanities point of view. This remains a proposed topic of further research using the approach and tooling presented.

[1] https://www.w3.org/standards/semanticweb/
[2] https://www.w3.org/TR/sparql11-query/
[4] https://vimeo.com/461293952

The paper is organized as follows. First, related work in epistolary historical network studies and temporal network analysis and systems are discussed to contextualize the work of this paper. Next a new data model and Linked Data sets conforming to it are presented as well as a LD service platform for publishing them, based on extending the traditional *5-star* model to a *7-star* model. After this, examples of network analyses using the Linked Data and SPARQL endpoint are presented. To test and demonstrate usability of the new data resource and data service even further, a semantic portal on top of the data service is presented with examples of data analyses. In conclusion, contributions of the paper and challenges of the proposed approach are summarized and discussed.

2. Related Work

2.1. Epistolary Historical Networks

During the Age of Enlightment it became suddenly possible for people to send and receive letters across Europe and beyond, based on a revolution in postal services, This opportunity resulted into what the contemporaries called the *Respublica litteraria*, Republic of Letters (RofL), a cross-national collaborative communication network that formed a basis for modern European scientific thinking, values, and institutions in Early Modern times 1400–1800. Data sources of early stage of Early Modern learned correspondences are proliferating rapidly, including, e.g., Europeana[5] [3], Kalliope Catalogue[6], The Catalogus Epistularum Neerlandicarum[7], Electronic Enlightenment[8], ePistolarium[9] [31], SKILLNET[10], correspSearch[11], the Mapping the Republic of Letters project[12], and Early Modern Letters Online (EMLO)[13] [10, 15, 27]. Visualizing the correspondences has been studied in the Mapping the Republic of Letters project[14] and in Tudor Networks of Power[15]. Bruneau et al. discuss applying Semantic Web Technologies to modelling the correspondences of French scientist Henri Poincaré and publishing on an online portal[16] [1].

The idea of representing epistolary data as a LD service was introduced in [36] and its application to DH research is discussed in [17] pointing out the analogy between RofL and Linked Open Data movement with some tooling, data analyses, and visualizations as examples. In this paper, the idea of using the Linked Data Service is developed and discussed further from a network analytic perspective, in relation to the correspondences in the four datasets listed in Table 1. We demonstrate flexibility and scientific potential of using an epistolary Linked Data Service for research in the following ways: (1) Firstly, by transforming and downloading the data into a suitable form, network analytic tools developed originally for different purposes, in our case for contemporary communication data, can be re-used, making it possible to apply them to historical epistolary networks, too. (2) Secondly, based on the Sampo model [19] and Sampo-UI framework [21], the data service can be integrated seamlessly with tooling for DH research making network analyses possible for researchers who often lack programming experience. (3) Thirdly, it is shown how the LD data service resource can be used for solving DH problems in network science with little programming experience using online programming services, such as Google CoLab[17] and Jupyter[18].

[5] http://www.europeana.eu
[6] http://kalliope.staatsbibliothek-berlin.de
[7] http://picarta.pica.nl/DB=3.23/
[8] http://www.e-enlightenment.com
[9] http://ckcc.huygens.knaw.nl/epistolarium/
[10] https://skillnet.nl
[11] https://correspsearch.net
[12] http://republicofletters.stanford.edu
[13] http://emlo.bodleian.ox.ac.uk
[14] http://republicofletters.stanford.edu/
[15] http://tudornetworks.net/
[16] http://henripoincare.fr/s/correspondance/page/accueil
[17] https://colab.research.google.com
[18] https://jupyter.org

2.2. Temporal Network Analysis

In the past few decades, communication data has become a relevant resource to understand the underlying social networks [29, 34]. In such cases, auto-recorded logs of pairwise interactions are modelled to construct a communication network, thus allowing the analysis of large-scale societal interactions and behavioural patterns. Here we focus on using epistolary Linked Data about communications to analyse historical correspondence networks of epistolary data but the methodology can equally well be used for modern communication networks, such as those from mobile phone logs, emails and social media platforms [37]. We identify two main approaches to analyzing such communication datasets according to the handling of temporality of the data [34, 37]. In a static approach a link is established between two people if there have been epistolary contacts between them, and in a temporal approach, the focus is on the distribution of dyadic interactions and behavioural features that characterize the way that people communicate. However, while most modern datasets attempt capture all auto-recorded communication within a communication channel (e.g., all emails or other communications within an organization [2, 5, 42]), this may not be true for historical data, since its collection is not automated, but implies broad manual compilation efforts by researchers.

For the static approach, a network is aggregated from dyadic interactions within a certain period or region. A link is created between two people if there has been some contact, and a proxy may be assigned for the strength of a tie based on, e.g., the total number of contacts [29]. From such static perspective it is possible to analyze large-scale properties of the resulting networks, including the degree distribution (i.e., the number of contacts of each node), different centrality measures (i.e., metrics to capture the relative importance of nodes within the network), or measures of the existence of communities or other types of structures.

For the temporal approach, a myriad of models have been proposed to analyzing network evolution [34]; we focus on the distribution of time-sequences of dyadic interactions, along with behavioural characteristics of how individual people communicate with their neighbours. From a sociological standpoint, the *Granovetter Effect* relates the notion of tie strength to network topology, noting that *strong ties* tend to be buried in overlapping circles of friends, akin to small communities where *weak ties* serve more as bridges between such communities [7, 29]. Since it is not possible to directly observe the strength of a tie, it is possible to use different temporal features as proxies [38]. Regarding the relationship of particular nodes to their neighbors, previous research [11, 33] has shown that individuals divide their contacting behaviour across their different neighbors in a persistent manner, known as a node's *social signature*, which is more stable in time than the neighbors themselves.

2.3. Using Linked Data for Network Analysis

The idea of using Linked Data graphs in network science is intuitive, natural, and not new. For example, in [8] linked data is transformed for network analysis for the LinkedDataLens system. In [30] RDF data is used for Social Network Analysis. Data from different sources can be aggregated into larger networks and enriched by each other and by inferring new triples, i.e., connections in the network. SPARQL queries and SPARQL CONSTRUCT can be used in flexible ways for network data transformations and creating tabular formats widely used. To facilitate network analysis and visualizations of RDF data there are tools available, such as the Semantic Web Import Plugin plugin[19] available for Gephi[20], arguably the leading visualization and exploration software for all kinds of graphs and networks. Applications of Gephi include, for example, Exploratory Data Analysis, Link Analysis, Social Network Analysis, and Biological Network analysis. A major contribution of our paper is to apply network analysis in a novel application domain for analysing historical epistolary communication networks, and especially by using temporal network analysis. For this purpose, a new LOD resource is presented and used.

[19] https://www.w3.org/2001/sw/wiki/GephiSemanticWebImportPlugin
[20] https://gephi.org/

3. A Linked Data Model and Service for Epistolary Data

This paper makes use of the epistolary datasets listed in Table 1. In our work these datasets were transformed into Linked Data and published according to the Linked Data publishing principles and other best practices of W3C [9], including, e.g., content negotiation and provision of a SPARQL endpoint. The CKCC corpus is to the best of our knowledge the first public linked open dataset on the Web on historical epistolary data; opening the publication of the correspSearch data in a similar way is done after getting a confirmation of the open license from the data owner.

3.1. Data Model for Linked Epistolary Data

By transforming the epistolary data into RDF we aimed to create knowledge graphs that include not only communication networks but also prosopographical data about the people and organizations involved. For this purpose a customised RDF-based metadata schema was created. The schema contains four different, interlinked classes: *Letter*, *Actor*, *Tie*, and *Place* as described in Table 2. Here the default namespace is the dataset-specific (*ckcc-schema*), *rdfs* refers to the RDF Schema[21], *crm* to the CIDOC CRM Schema[22], *geo* to WGS84 Geo Positioning vocabulary[23], *skos* to SKOS Simple Knowledge Organization System namespace[24], and *xsd* to the XML Schema of W3C[25].

The design choices are based on the principles developed in the EMLO project [36]. In the epistolary dataset, instances of the class *Actor* can be either people or groups. Each actor is connected to the sent letters using the property *:created*. Each letter is modeled as an instance of the class *Letter* that has seven properties describing the letter. A letter is linked with its recipients using the property *:was_addressed_to*, to places of sending and receiving using the properties *:was_sent_from* and *:was_sent_to*, and to related timespan with *crm:P4_has_time-span*. Furthermore, a letter instance is enriched with information about the data source and a human-readable description. The correspondences between two actors are modeled as instances of the class *Tie*. Each of these instances is linked to the two actors and likewise each letter is linked to the corresponding tie. Using the *Tie* instance simplifies the database queries e.g. in cases of querying all the letters between the two actors. In addition, this model facilitates to adding precalculated network metrics such as node degrees and centrality measures to the data model. In addition, the data set also contains precalculated values for the time of flourishing for each actor, e.g. the time period when the actor has been active in letters correspondences. The resources in the domain ontology of the places consist of place labels, the coordinate information, and the hierarchy built with the property *skos:broader*. Finally, the timespans follow the four point model, e.g. with *xsd:dateTime* values indicating the earliest and latest moments for the beginning and the end.

The two datasets, CKCC and correspSearch, were converted from different source formats. CKCC is an extract from an existing RDF dataset, while the correspSearch data was converted from a source in JSON format. In these datasets both the actor and place resources had linkage to external LOD cloud databases, e.g., Wikidata, VIAF, Early Modern Letters Online project (EMLO), or database of Deutsche Nationalbibliothek[26] (GND). This existing linkage was used for two main purposes. First, in the current data publication, the resources in the datasets where reconciled based on the links, e.g. the actors or places refer to the same entity, if they point to the same external link. Secondly, the external databases were used to enrich our data e.g. with images of actors and coordinates of the places. In our work, the "FAIR[27] guiding principles for scientific data management and stewardship" of publishing data are used[28].

[21] https://www.w3.org/TR/rdf-schema/
[22] http://www.cidoc-crm.org
[23] http://www.w3.org/2003/01/geo/wgs84_pos#>
[24] http://www.w3.org/2004/02/skos/core#
[25] https://www.w3.org/XML/Schema
[26] https://www.dnb.de/EN/Home/home_node.html
[27] FAIR: Findable, Accessible, Interoperable, and Re-usable
[28] https://www.go-fair.org/fair-principles/

Table 2

RDF schema for Letter, Actor, Tie, and Place. Column *C* marks the cardinality of the element. Fields inferred from the data are marked with *cursive* text.

Element URL	C	Range	Meaning of the value
ACTOR			
skos:prefLabel	1	xsd:string	Preferable label
:created	0..n	:Letter	Created letter
:birthDate	0..1	crm:E52_Time-Span	Time of birth
:birthPlace	0..1	crm:E53_Place	Place of birth
:flourished	0..1	crm:E52_Time-Span	*Time of flourishing*
:deathDate	0..1	crm:E52_Time-Span	Time of death
:deathPlace	0..1	crm:E53_Place	Place of death
:has_statistic	1...n	:NetworkStatistic	*Precalculated network statistics, e.g., centrality measures*
:source	1..n	rdfs:Resource	Used data source
LETTER			
skos:prefLabel	1	xsd:string	Preferable label
:was_addressed_to	0..1	crm:E39_Actor	Recipient of the letter
:was_sent_from	0..1	crm:E53_Place	Place of sending
:was_sent_to	0..1	crm:E53_Place	Place of receiving
crm:P4_has_time-span	0..1	crm:E52_Time-Span	Time of sending
:source	1..n	rdfs:Resource	Used data source
:in_tie	1	:Tie	*Correspondence in which this letter belongs to*
TIE			
:actor1	1	crm:E39_Actor	First correspondent
:actor2	1	crm:E39_Actor	Second correspondent
:num_letters	1	xsd:integer	*Number of letters in this correspondence*
skos:prefLabel	1	xsd:string	Preferable label
PLACE			
crm:P89_falls_within	0..1	crm:E53_Place	Place higher in hierarchy
skos:prefLabel	1	xsd:string	Preferable label
geo:lat	0..1	xsd:decimal	Latitude of the coordinates
geo:long	0..1	xsd:decimal	Longitude of the coordinates
TIMESPAN			
crm:P82a_begin_of_the_begin	0..1	xsd:dateTime	Earliest time for the beginning
crm:P81a_end_of_the_begin	0..1	xsd:dateTime	Latest time for the beginning
crm:P81b_begin_of_the_end	0..1	xsd:dateTime	Earliest time for the end
crm:P82b_end_of_the_end	0..1	xsd:dateTime	Latest time for the end
skos:prefLabel	1	xsd:string	Preferable label

3.2. Using the Linked Data and Data Service

The data can be used for research via (1) ready-to-use tools available on a semantic portal or (2) by using the underlying SPARQL endpoint with external tools, based on a framework called *LetterSampo* [17]. The SPARQL endpoint can be used directly in DH research using, e.g., YASGUI[29] [32] and Python scripting in Google Colab or Jupyter notebooks. The endpoint can also be used for filtering and downloading the data in different forms, such as in tabular CSV format, for external data-analysis tools, in our case for network analyses.

This framework is used for creating data services and semantic portals[30] based on the Sampo model [19] for sharing collaboratively enriched linked open data using a shared ontology infrastructure. The portals host ready-to-use data-analytic tools for DH research, as suggested in [18]. The Sampo-UI framework [21] is used as the interface model and as the full stack JavaScript tool. Sampo portals are based—from a data perspective—on querying the SPARQL endpoint from the client side using JavaScript. The portals in the Sampo series demonstrate the idea that versatile web applications can be implemented by separating the application logic and data services via SPARQL API, which arguably facilitates developing new applications efficiently by re-using the same data.

[29] https://yasgui.triply.cc

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

3.3. Querying and rendering networks in a web portal

The networks in the portal pages are constructed using a customizable back-end service Sparql2GraphServer [25]. It was developed to meet the requirements for querying and constructing a network from any SPARQL endpoint. It builds a Sampo-UI compatible network based on SPARQL queries. It is a Python application built on Flask[31] framework using modules SPARQLWrapper[32] and Networkx. The visual appearance of the network on a portal page is configured in the front-end Sampo-UI settings. The back-end service is used in other portals in the Sampo series like AcademySampo and ParliamentSampo. Figure 1 depicts a network extracted from Wikidata, it illustrates the teacher–student relationships starting from German polymath Gottfried Wilhelm Leibniz.

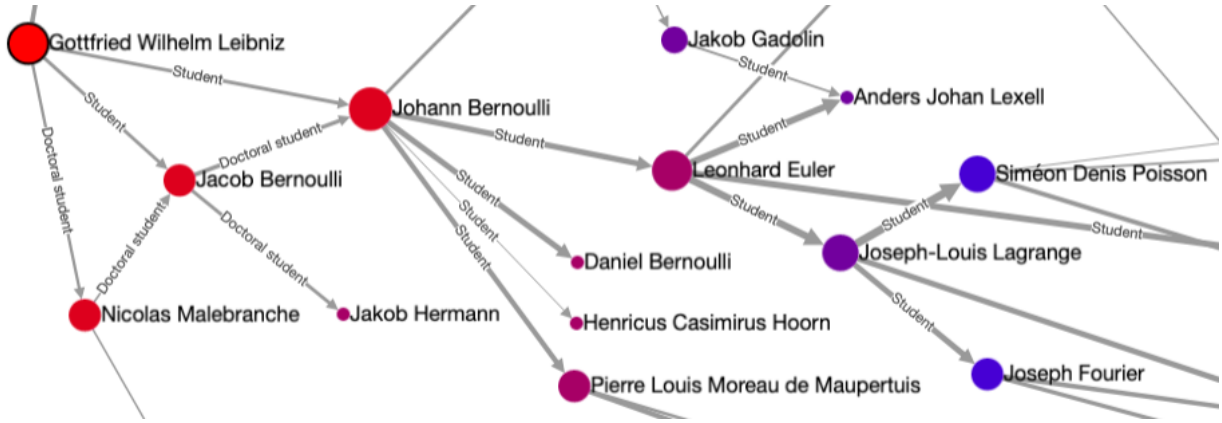


Figure 1. Social network of polymath Gottfried Wilhelm Leibniz in Wikidata

3.4. New Resources on the Web

The CKCC knowledge graph[33] as well as the correspSearch knowledge graph[34] have been published on the Linked Data Finland platform LDF.fi [20]. Both dataset are also available at Zenodo[35]. LDF.fi uses the *7-star* model for LD deployment [20] that extends the *5-star* model[36] coined by Tim Berners-Lee: to enhance re-usability of LD, the *sixth star* is given if the data is published with its schema and the *seventh star* if validation results of the data using the schema are provided. LDF.fi is powered by the Fuseki SPARQL server[37] and Varnish Cache web application accelerator[38] for routing URIs, content negotiation, and caching. The portal user interface was implemented by the Sampo-UI framework [21]. The system uses Docker microservice architecture containers[39]. By using containers, the services can be migrated to another computing environment easily, and third parties can re-use and run the services on their own. The architecture also allows for horizontal scaling for high availability, by starting new container replicas on demand. The framework will be used in the *Constellations of Correspondence (CoCo)* project[40] on correspondences in the Grand Duchy of Finland in the 19th century [35].

[31] https://flask.palletsprojects.com/en/2.2.x/
[32] https://sparqlwrapper.readthedocs.io
[33] The data, schema, and service are openly available (CC BY-NC 4.0) at the homepage https://www.ldf.fi/dataset/ckcc.
[34] The data, schema, and service available (CC BY-NC 4.0) at https://www.ldf.fi/dataset/corresp.
[35] CKCC: https://zenodo.org/record/6631385, correspSearch: https://zenodo.org/record/5972316
[36] https://5stardata.info/en/
[37] https://jena.apache.org/documentation/fuseki2/
[38] https://varnish-cache.org
[39] https://www.docker.com
[40] https://seco.cs.aalto.fi/projects/coco/

4. Network Analyses Using the Linked Data Service

In this section we first show some general network analyses results of the epistolary datasets of Table 1. After this, it is shown how the SPARQL endpoint can be used for research using querying and by programming. For these purposes, examples using the data with custom network analytic tools, Yasgui and Google Colab are presented, respectively. Finally, analyzing the data with ready-to-use tools and the two-step analysis model of the LetterSampo portal is discussed with examples.

4.1. Exporting Data for Data Analyses

A simple way of reusing the data resources is to download and transfer them for the analysis tool of choice. For this purpose either data dumps from Zenodo or the SPARQL endpoint can be used. A benefit of using the endpoint is that the data can be filtered and even transformed during the download to fit better for the aimed purpose. An example of using the data resource in external network analytic tools is presented in [39]. In this case study, the Linked Data of CKCC and correspSearch were analyzed in terms of network metrics and compared with four modern datasets of mobile phone call networks, emails, community boards, and wall-postings on a social media platform. It turned out that contemporary and historical epistolary communication networks resemble each other strikingly even if the media were quite different.

4.2. General Analyses on Epistolary Networks

The network portal also shows the precalculated centrality measures for each actor. First, a correspondence network was created from the RDF data and thereafter the measures where calculated using the Python library NetworkX[41]. These measures are based on a network containing both the CKCC and the correspSearch datasets.

Table 3
Precalculated network measures for René Descartes

Measure	**Value**	**Rank**
Betweenness Centrality	0.00930	6
Clique Number	4	14
Clustering Coefficient	0.000162	380
Core Number	7	1
Eigenvector Centrality	0.064	5
Number of Correspondences	92	12
Pagerank Centrality	0.00417	23
Weighted In-Degree	164	16
Weighted Out-Degree	585	5

An example of the measures for French philosopher and scientist René Descartes are listed in Table 3. In the table, e.g., the *Clique Number* with a value of 4 indicates that Descartes is a part of complete subgraph where all the nodes have a degree of 4, and the rank of 14 indicates that there are 13 larger cliques in the entire network. The *Weighted Out-* and *In-Degrees* correspond to the total number of sent and received letters while the *Number of Correspondences* equals the unweighted node degree. Also the actor perspective facet page has a socio-centric network visualization where the actors can be filtered e.g. by their gender, years of living, or data sources.

[41] https://networkx.org

```
PREFIX lssc: <http://ldf.fi/schema/lssc/>
PREFIX crm: <http://www.cidoc-crm.org/cidoc-crm/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT DISTINCT ?source_label ?target_label ?letter_label ?date
WHERE {
  VALUES ?source { <http://ldf.fi/ckcc/actors/p300075> }

  ?source lssc:created ?letter ;
          skos:prefLabel ?source_label .

  ?letter a lssc:Letter ;
          lssc:was_addressed_to ?target ;
          skos:prefLabel ?letter_label ;
          lssc:has_time/skos:prefLabel ?date .

  ?target skos:prefLabel ?target_label .

} ORDER BY ?date
```

Figure 2. SPARQL example for querying the letters by René Descartes

4.3. Querying the SPARQL Endpoint

For the analyses presented in this article, there are basically two practices for using a SPARQL endpoint. Firstly, for showing the data results on the web portal, the tabular results of a relatively simple query are shown on the portal page. An example of such of query is shown in Fig 2. It queries all letters sent by Descartes and shows their recipients, labels, and dates sorted by the date. Secondly, analyzing or visualizing network structures may require several database queries e.g. for separated lists of actors (*nodes*) and letters (*edges*). The actual results are thereafter calculated based on the data of these simple, straight-forward queries with spreadsheet-like results.

4.4. Using the LetterSampo Portal

Also a portal demonstrator[42] based on the aggregated CKCC and correspSearch LOD was published on the Web for public use. [17] The network portal provides components for visualizing the epistolary data using line charts, maps, and networks. Figure 3 depicts an egocentric network around Descartes. In this visualization the widths of the edges are proportional to the number of letters between the two actors while the sizes of the nodes are based on the length of the shortest path between the nodes so that the main actor appears with the largest node and the most distant actors have the smallest nodes. In spite that Descartes is the center actor, Constantijn Huygens has a higher node degree due to the fact that the CKCC dataset contains a larger amount of letters by him.

Figure 4 depicts a visualization of the *social signatures* [11, 33] of Descartes. Social signatures represent how individuals communicate with their neighbours in a given time. This visualization has curves for his entire time of flourishing (blue line) and separated curves during his career, e.g. the red line for time period 1643–1650. For an interval (e.g., 1631–1637, 1637–1643), a social signature is obtained by (1) computing the fraction of outgoing contacts per alter, and (2) ranking the alters. In the chart, like for instance the highest value of the yellow line is 0.368 indicating that Descartes wrote 36.8% of his correspondences to the top ranked alter, and likewise 33.3% to the second alter. This approach allows characterizing the relative importance of different alters in an ego network. When comparing different individuals, it is found that their social signatures tend to be stable [11, 33, 37].

4.5. Using the Endpoint by Programming

Due to the performance issues when attempting to render a larger network of more than, e.g., 1000 nodes on a browser page, data was further visualized in Google Colabs environment using Python. As an example, the largest

[42] https://lettersampo.demo.seco.cs.aalto.fi/en

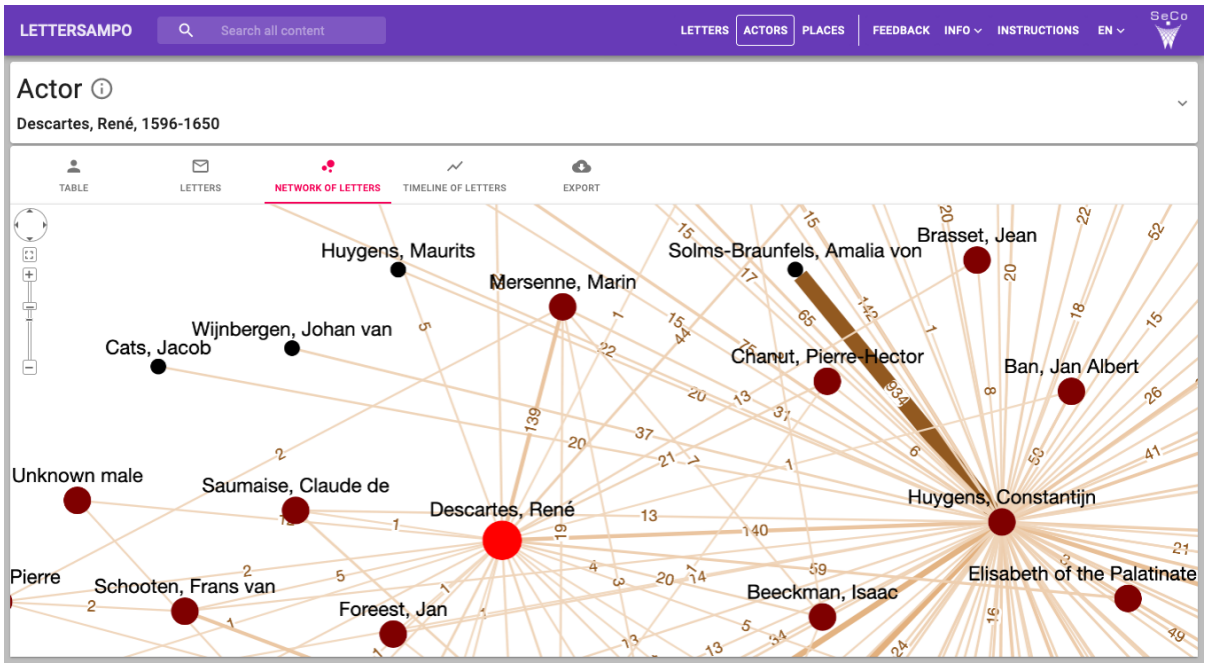


Figure 3. Network of correspondences around René Descartes

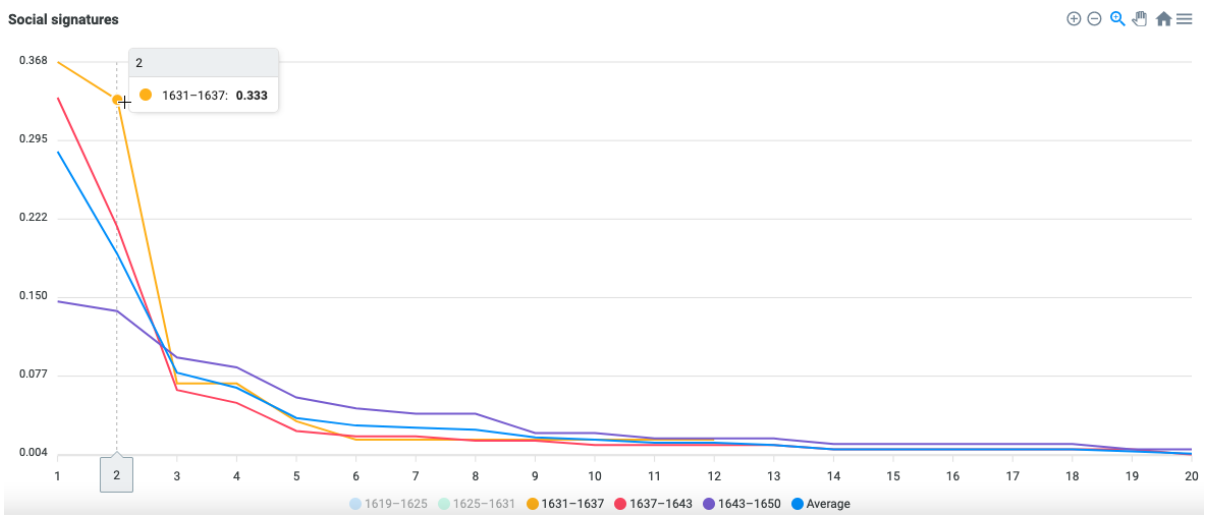


Figure 4. Chart depicting the social signatures of René Descartes

connected component of the CKCC data is visualized in Figure 5. The network is built around three center actors: Dutch poet and composer Constantijn Huygens, philosopher Hugo de Groot, and mathematician and physicist Christiaan Huygens, who have high node degree values. On the other hand, there is a multitude of actors with low node degree. As a comparison the correspSearch data in Figure 6 has much more of these hubs.

Figure 7 depicts the correspondences of Descartes on a timeline. The entire timeline is shown on the lower part of the chart. On the upper part of the chart there are separately the ten most active correspondences of Descartes and the lowest line depicts the correspondences with all the other actors. The visualization also reveals biases

Figure 5. Network of CKCC data

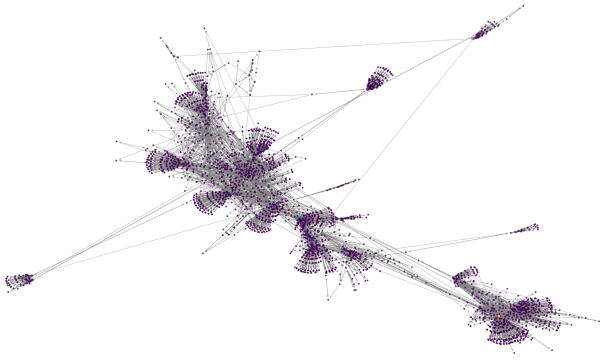


Figure 6. Network of correspSearch data

caused by missing information in the source data. For example, when studying the correspondence with French philosopher and mathematician Marin Mersenne, it can be observed that the source collections contains 134 letters from Descartes to Mersenne, but only of five by Mersenne to Descartes.

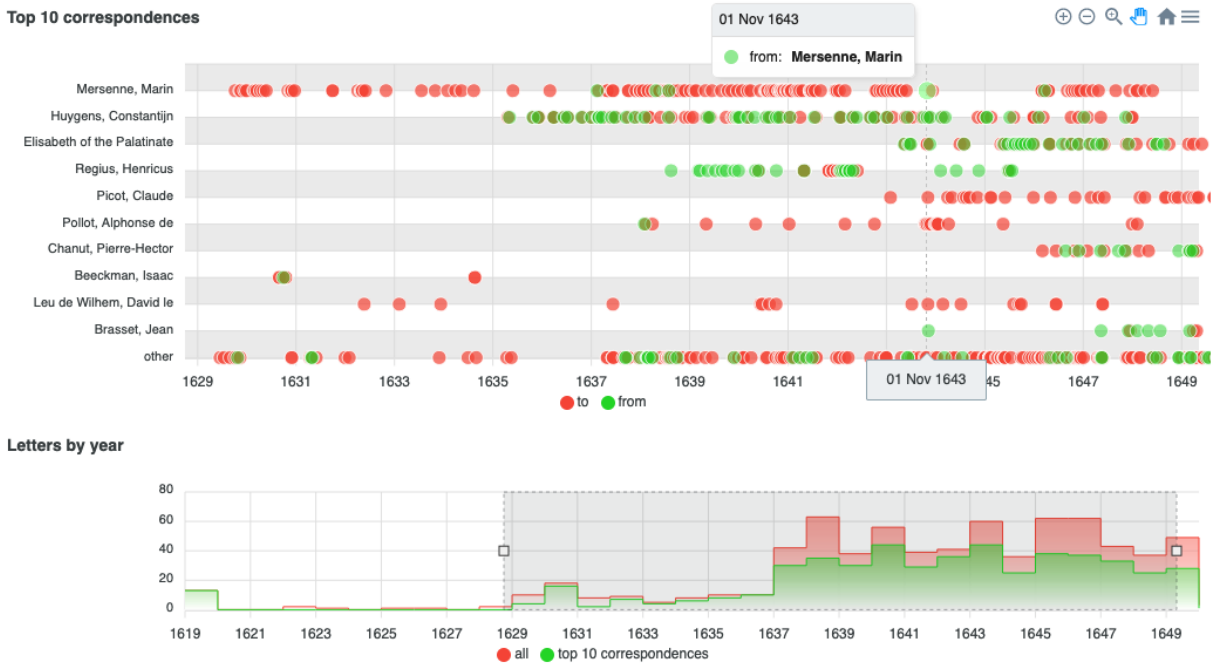


Figure 7. Timeline depicting top 10 letter correspondences of René Descartes

Figure 8 depicts the most active scientists by the decades 1620–1690. The ranking is based on the total amount of sent and received letters and the data is visualized so that the first ranking scientist is on the top of the chart. The figure depicts that from 1620 to 1640 Descartes is on the first rank, but later replaced by Christiaan Huygens. The code is available in GitHub[43] including a link to notebook in Google Colaboratory.

[43]https://github.com/SemanticComputing/LetterSampo-ckcc-charts

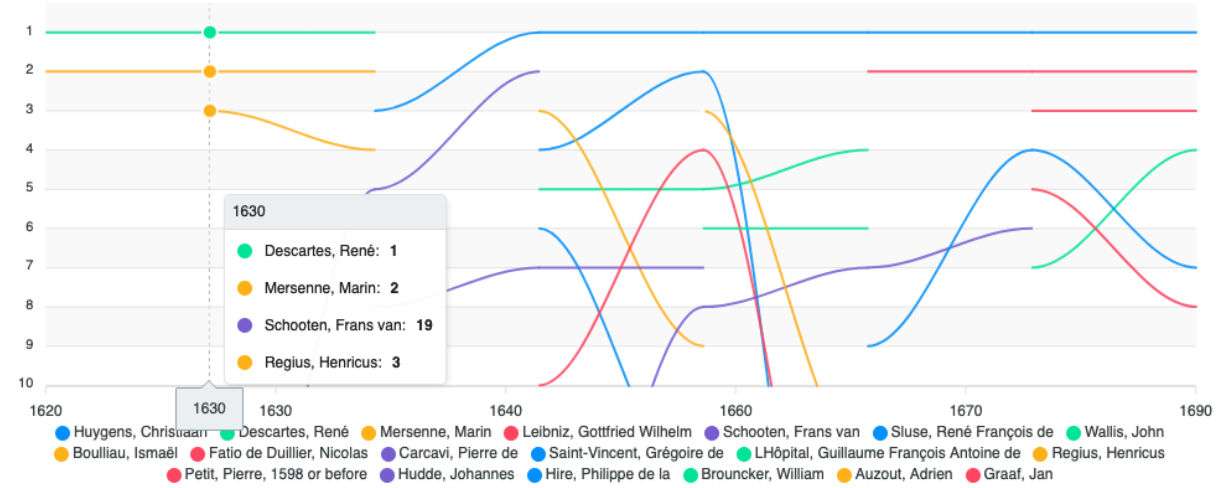


Figure 8. Top scientists in the CKCC data during 1620–1690

4.6. *Comparing Contemporary and Historical Communication Networks*

As a use-case scenario, a comparative analysis of historical and contemporary communication networks was performed, with results formally introduced in [37]. In this study, the goal was to compare aspects of temporal communication networks at different granularity levels, including snapshots of static graphs, the time series of dyadic interactions, as well as ego networks. We compare these features with different contemporary communication channels, including emails, social media platforms, forums and mobile phone calls.

In brief, the goal was to analyze to what extent different behavioural features of contemporary communication networks can be found in historical datasets. We find similarities at different degrees of success. Particularly, we find evidence for the persistence of social signatures in historical context, as well as the Granovetter effect for different proxies of tie strength, and important similarities in the distribution of dyadic timings. We found, however, difficulties in drawing conclusions from global network analyses, particularly given that some individuals are over-represented in historical datasets and the data is biased.

Regarding social signatures, the results suggest that individuals divide their communication similarly across top-ranked alters; in other words, that the social signatures of a given individual are more similar among different periods than to the signatures of different egos. These results were consistent across different filters for the constructions of ego networks. Taken together, they suggest that in practice individuals allocate time and resources systematically when communicating. Regarding the Granovetter effect, it is found that stronger ties are associated with overlapping circles of friends, a feature that persists even when considering different proxies for the strength of ties. While these results have been previously observed in contemporary datasets [11, 29, 33, 38], historical datasets bring an added value on two fronts: (1) They provide evidence for human communication patterns in a distinct context—contemporary examples are usually the result of auto-recorded digital logs, and are thus representative of modern practices. (2) They provide a timeframe that is unachievable in contemporary datasets, where samples of ego networks are examined across different decades, and where aggregate network evolution spans centuries.

5. Discussion

This paper presented tools and systems for analyzing networks of epistolary linked open data. Of the four datasets discussed, CKCC and correspSearch datasets are, to the best of our knowledge, first LOD-based epistolary datasets available on the Semantic Web. Examples of analyzing and visualizing the network data were presented and discussed using SPARQL querying and Python scripting as a proof-of-concept of the usability of the data resources.

The aggregated data of these two datasets are openly available for the research community for related analyses. We also demonstrated the idea of developing applications, i.e., semantic portals, on top of the data service that require no programming skills from the end user.

This paper focused on presenting, discussing, and illustrating design principles for publishing and using epistolary network data as Linked Data, not on presenting actual analysis results of particular datasets. This remains a topic of further research, but the first experiments presented show in our opinion that the framework and the published resources, the linked open data and data service at LDF.fi, and the LetterSampo portals are promising in filtering our patterns of possibly interesting phenomena in Big Data using distant reading [28]. However, traditional close reading by a human is needed as before in interpreting the results.

A major challenge in creating data analyses like the ones shown in this paper is related to the quality of the data produced. Historical (meta)data is typically incomplete and our knowledge about it is uncertain. Also using more or less automatic means for transforming and linking the data leads to problems of incomplete, skewed, and erroneous data [26]. In historical epistolary data in particular, the data is seldom complete as only part of the letters have survived or are included in the data available. The data is often also biased in different ways because historical data is often a result of a collection process performed by humans. For example, only letters of significant people have typically been collected in archives. It is therefore difficult to compare the underlying network with some modern networks, such as mobile phone networks, where the data has not been subject to human selection and is complete. This problem could be addressed by collecting data in unbiased ways or by trying analyze afterwards in what ways the data is biased.

This as well as conceptual difficulties in modeling complex real world ontologies, such as historical geogazetteers, become sometimes embarrassingly visible when using and exposing the knowledge structures to end-users. The same problems exist in traditional systems but are hidden in the non-structured presentations of the data. In general, more data literacy [24] is usually needed from the end-user when using data analytic tools.

The methods of network analysis can be very sensitive to even small errors in the data or biases in the sampling schemes. For example, the values of betweenness centrality can dramatically change by removal of even a single link, or long silences in communication in historical data can be explained by missing data from some historical period rather than inherently bursty communication tendencies. While computing various measures based on network data can be relatively simple with tools that are introduced here, the remaining challenge is to correctly interpret the results. This requires expert knowledge both in the domain to know how the data is biased and the methods to know how this affects the various measures. In the future, sampling schemes and missing data could be encoded in the data framework and the measures could be adopted to handle these situations. However, this work would first need to be done within the domains (e.g., encoding sampling details of historical correspondence) and network method development (e.g., measures that consider missing data [23]).

The datasets CKCC and correspSearch contained linkage to external LOD cloud databases which facilitated enriching the data by extracting, e.g., information about the lifespans of the actors or geological metadata of places. Communication networks are easily huge, consisting of millions of links, which causes performance issues when, e.g., querying the database or rendering a large network on the web portal.

In spite of the challenges inherent in historical epistolary data, application of network analysis to the data can be useful for the researchers in finding out potentially interesting patterns of knowledge for closer study in datasets that are too big or complex for traditional manual means only. The new LOD resources and applications presented in this paper can now be used for this purpose.

Acknowledgements

We thank Arno Bosse, Howard Hotson, and Miranda Lewis for earlier collaborations during the *Cultures of Knowledge* project at the University of Oxford, funded by the Mellon Foundation, Charles van den Heuvel and Dirk van Miert for discussions related to CKCC, as well as colleagues in the EU COST Action project *Reassembling the Republic of Letters*[44]. Stefan Dumont provided the correspSearch data for our use. The work was also part of

[44] http://www.republicofletters.net

the *Open Science and Research Programme*[45], funded by the Ministry of Education and Culture of Finland, and the EU project InTaVia: In/Tangible European Heritage[46], and is related to the EU COST action Nexus Linguarum[47] on linguistic data science. CSC – IT Center for Science[48] provided computational resources.

References

[1] Bruneau, O., Lasolle, N., Lieber, J., Nauer, E., Pavlova, S., Rollet, L.: Applying and developing semantic web technologies for exploiting a corpus in history of science: The case study of the Henri Poincaré correspondence. Semantic Web **12**(2), 359–378 (2021)

[2] Diesner, J., Frantz, T.L., Carley, K.M.: Communication networks from the Enron email corpus "It's always about the people. Enron is no different". Computational & Mathematical Organization Theory **11**(3), 201–228 (2005). , https://doi.org/10.1007/s10588-005-5377-0

[3] Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., Van de Sompel, H.: The europeana data model (edm). In: World Library and Information Congress: 76th IFLA general conference and assembly. vol. 10, p. 15. IFLA (2010)

[4] Dumont, S.: correspSearch -– Connecting Scholarly Editions of Letters. Journal of the Text Encoding Initiative (10) (2016). , https://doi.org/10.4000/jtei.1742

[5] Eckmann, J.P., Moses, E., Sergi, D.: Entropy of dialogues creates coherent structures in e-mail traffic. Proceedings of the National Academy of Sciences **101**(40), 14333–14337 (2004), https://doi.org/10.1073/pnas.0405728101

[6] Goh, K.I., Barabási, A.L.: Burstiness and memory in complex systems. EPL (Europhysics Letters) **81**(4), 48002 (Jan 2008). , https://doi.org/10.1209/0295-5075/81/48002

[7] Granovetter, M.S.: The Strength of Weak Ties. American Journal of Sociology **78**(6), 1360–1380 (1973). , https://doi.org/10.1086/225469

[8] Groth, P., Gil, Y.: Linked data for network science. In: Proceedings of the First International Conference on Linked Science - Volume 783. pp. 1–12. LISC'11, CEUR-WS.org, Aachen, DEU (2011)

[9] Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool (2011), http://linkeddatabook.com/editions/1.0/

[10] van den Heuvel, C.: Mapping knowledge exchange in Early Modern Europe: Intellectual and technological geographies and network representations. International Journal of Humanities and Arts Computing **9**(1), 95–114 (3 2015). , https://doi.org/10.3366/ijhac.2015.0140

[11] Heydari, S., Roberts, S.G., Dunbar, R.I.M., Saramäki, J.: Multichannel social signatures and persistent features of ego networks. Applied Network Science **3**(1) (May 2018). , https://doi.org/10.1007/s41109-018-0065-4

[12] Hitzler, P.: A Review of the Semantic Web Field. Commun. ACM **64**(2), 76–83 (Jan 2021). , https://doi.org/10.1145/3397512

[13] Hitzler, P., Krötzsch, M., Rudolph, S.: Foundations of Semantic Web technologies. Springer–Verlag (2010), https://doi.org/10.1201/9781420090512-17

[14] Holme, P., Saramäki, J. (eds.): Temporal Network Theory. Springer–Verlag (2019), https://doi.org/10.1007/978-3-030-23495-9

[15] Hotson, H., Wallnig, T. (eds.): Reassembling the Republic of Letters in the Digital Age. Göttingen University Press (2019), https://doi.org/10.17875/gup2019-1146

[16] Hyvönen, E.: Publishing and using cultural heritage linked data on the semantic web. Morgan & Claypool, Palo Alto, CA (2012), https://doi.org/10.2200/S00452ED1V01Y201210WBE003

[17] Hyvönen, E., Leskinen, P., Tuominen, J.: LetterSampo – Historical Letters on the Semantic Web: A Framework and Its Application to Publishing and Using Epistolary Data of the Republic of Letters. Journal on Computing and Cultural Heritage **16**(1) (2023), https://doi.org/10.1145/3569372

[18] Hyvönen, E.: Using the semantic web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. Semantic Web – Interoperability, Usability, Applicability **11**(1), 187–193 (2020), https://doi.org/10.3233/SW-190386

[19] Hyvönen, E.: Digital humanities on the semantic web: Sampo model and portal series. Semantic Web – Interoperability, Usability, Applicability pp. 1–16 (2022), https://doi.org/10.3233/SW-223034

[20] Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: Proceedings of the ESWC 2014 Demo and Poster Papers. Springer–Verlag (2014), https://doi.org/10.1007/978-3-319-11955-7_24

[21] Ikkala, E., Hyvönen, E., Rantala, H., Koho, M.: Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. Semantic Web – Interoperability, Usability, Applicability **13**(1), 69–84 (January 2022), https://doi.org/10.3233/SW-210428, online version published in 2021, print version in 2022

[22] Karsai, M., Kivelä, M., Pan, R.K., Kaski, K., Kertész, J., Barabási, A.L., Saramäki, J.: Small but slow world: How network topology and burstiness slow down spreading. Physical Review E **83**(2) (Feb 2011), https://doi.org/10.1103/physreve.83.025102

[23] Kivelä, M., Porter, M.A.: Estimating interevent time distributions from finite observation periods in communication networks. Physical Review E **92**(5), 052813 (2015), https://doi.org/10.1103/physreve.92.052813

[45] http://openscience.fi

[46] https://intavia.eu/

[47] https://nexuslinguarum.eu/the-action

[48] https://www.csc.fi/en/home

[24] Koltay, T.: Data literacy for researchers and data librarians. Journal of Librarianship and Information Science **49**(1), 3–14 (2015). , https://doi.org/10.1177/0961000615616450
[25] Leskinen, P., Hyvönen, E., Tuominen, J.: Sparql2GraphServer: a Server-side Tool for Extracting Networks from Linked Data for Data Analysis. In: ISWC-Posters-Demos-Industry 2021 International Semantic Web Conference (ISWC) 2021: Posters, Demos, and Industry Tracks. CEUR Workshop Proceedings (Oct 2021), http://ceur-ws.org/Vol-2980/paper343.pdf
[26] Mäkelä, E., Lagus, K., Lahti, L., Säily, T., Tolonen, M., Hämäläinen, M., Kaislaniemi, S., Nevalainen, T.: Wrangling with non-standard data. In: Proceedings of the Digital Humanities in the Nordic Countries 5th Conference. pp. 81–96. CEUR Workshop Proceedings (2020), http://ceur-ws.org/Vol-2612/paper6.pdf
[27] van Miert, D.: What was the Republic of Letters? A brief introduction to a long history (1417–2008). Groniek **204/205**, 269–287 (2016)
[28] Moretti, F.: Distant Reading. Verso Books (2013), https://doi.org/10.1093/llc/fqu010
[29] Onnela, J.P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., Barabási, A.L.: Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences **104**(18), 7332–7336 (Apr 2007), https://doi.org/10.1073/pnas.0610245104
[30] Raji, P.S., Surendran, S.: RDF approach on social network analysis. In: 2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS). pp. 1–4 (2016). , https://doi.org/10.1109/rains.2016.7764416
[31] Ravenek, W., van den Heuvel, C., Gerritsen, G.: The epistolarium: origins and techniques. CLARIN in the Low Countries pp. 317–323 (2017), https://doi.org/10.5334/bbi.26
[32] Rietveld, L., Hoekstra, R.: The YASGUI family of SPARQL clients. Semantic Web **8**(3), 373–383 (2017), https://doi.org/10.3233/sw-150197
[33] Saramaki, J., Leicht, E.A., Lopez, E., Roberts, S.G.B., Reed-Tsochas, F., Dunbar, R.I.M.: Persistence of social signatures in human communication. Proceedings of the National Academy of Sciences **111**(3), 942–947 (Jan 2014). , https://doi.org/10.1073/pnas.1308540110
[34] Saramäki, J., Moro, E.: From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. The European Physical Journal B **88**(6) (2015). , https://doi.org/10.1140/epjb/e2015-60106-6
[35] Tuominen, J., Koho, M., Pikkanen, I., Drobac, S., Enqvist, J., Hyvönen, E., Mela, M.L., Leskinen, P., Paloposki, H.L., Rantala, H.: Constellations of Correspondence: a Linked Data Service and Portal for Studying Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland. In: DHNB 2022 The 6th Digital Humanities in Nordic and Baltic Countries Conference. pp. 415–423. CEUR Workshop Proceedings, Vol. 3232 (March 2022), http://ceur-ws.org/Vol-3232/paper41.pdf
[36] Tuominen, J., Mäkelä, E., Hyvönen, E., Bosse, A., Lewis, M., 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). pp. 76–88. CEUR Workshop Proceedings, vol. 2084 (March 2018), http://www.ceur-ws.org/Vol-2084/paper6.pdf
[37] Ureña-Carrion, J., Leskinen, P., Tuominen, J., Hyvönen, E., Kivelä, M.: Communication now and then: Analyzing the Republic of Letters as a communication network. Applied Network Science **7**(26) (2022), https://doi.org/10.1007/s41109-022-00463-1
[38] Ureña-Carrion, J., Saramäki, J., Kivelä, M.: Estimating tie strength in social networks using temporal communication data. EPJ Data Science **9**(1) (Dec 2020), https://doi.org/10.1140/epjds/s13688-020-00256-5
[39] Ureña-Carrion, J., Leskinen, P., Tuominen, J., van den Heuvel, C., Hyvönen, E., Kivelä, M.: Communications now and then: Analyzing the Republic of Letters as a communication network. Applied Network Science (2022), https://arxiv.org/abs/2112.04336v1, in press
[40] Vespignani, A.: Twenty years of network science. Nature **558**(7711), 528–529 (Jun 2018), https://doi.org/10.1038/d41586-018-05444-y
[41] Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**(6684), 440–442 (Jun 1998). , https://doi.org/10.1038/30918
[42] Wu, Y., Zhou, C., Xiao, J., Kurths, J., Schellnhuber, H.J.: Evidence for a bimodal distribution in human communication. Proceedings of the National Academy of Sciences **107**(44), 18803–18808 (2010), https://doi.org/10.1073/pnas.1013140107

Publication XII

Petri Leskinen, Eero Hyvönen, and Jouni Tuominen. Members of Parliament in Finland Knowledge Graph and its Linked Open Data Service. *Further with Knowledge Graphs. Proceedings of the 17th International Conference on Semantic Systems, 6-9 September 2021, Amsterdam, The Netherlands.*, Mehwish Alam, Paul Groth, Victor de Boer, Tassilo Pellegrini, Harshvardhan J. Pandit, Elena Montiel, Víctor Rodríguez Doncel, Barbara McGillivray, Albert Meroño-Peñuela (editors), IOS Press, pages 255–269, DOI 10.3233/SSW210049, online https://doi.org/10.3233/SW-210049 .

Members of Parliament in Finland Knowledge Graph and its Linked Open Data Service

Petri Leskinen[1][0000−0003−2327−6942], Eero Hyvönen[1,2][0000−0003−1695−5840], and Jouni Tuominen[1,2][0000−0003−4789−5676]

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

Abstract. This paper presents a prosopographical knowledge graph describing the Members of Parliament in Finland and related actors in politics, extracted from the databases and textual descriptions of the Parliament of Finland. The data has been interlinked internally and enriched with data linking to external data sources according to the 5-star Linked Data model. The data has been published together with its schema for better re-usability and is validated using ShEx. The knowledge graph presented is integrated with another knowledge graph about over 900 000 parliamentary plenary speeches in Finland (1907–) to form a larger parliamentary LOD publication *FinnParla* of the Parliament of Finland. The data is being used for Digital Humanities research on parliamentary networks, culture, and language.

Keywords: Parliamentary data · Biographies · Linked Data · Digital Humanities · Entity linking

1 Introduction

A key idea of Linked Data [9] is to enrich datasets by integrating complementary local information sources in an interoperable way into a global knowledge graphs [8] to be used in applications. This involves harmonization of the local data models used, as well as aligning the concepts and entities (resources) used in populating the local data models.

This paper reports first results of the Semantic Parliament (SEMPARL)[3] project that produces a Linked Open Data (LOD) and research infrastructure for Finnish parliamentary data, and develops novel semantic computing technologies and applications to study parliamentary political culture and language. The project is related to various similar efforts in other countries [2,20,5] and in EU [1]; parliamentary open data is an important asset for rendering political decision making transparent, and such data is widely used for research on political language and culture.

SEMPARL aims at three major contributions:

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

1. The project responds to the demand for an easy to use and "intelligent" access to the newly digitized Finnish parliamentary data by providing the data as a national Linked Open Data (LOD) infrastructure and service for researchers, citizens, government, media, and application developers.
2. The project studies long-term changes in the Finnish parliamentary and political culture and language. These use cases are pioneering studies using the Finnish digital parliamentary data.
3. The new LOD service enriches semantically content in other related Finnish LOD services, such as LawSampo for Finnish legislation and case law [14] and BiographySampo [13] for prosopographical data.

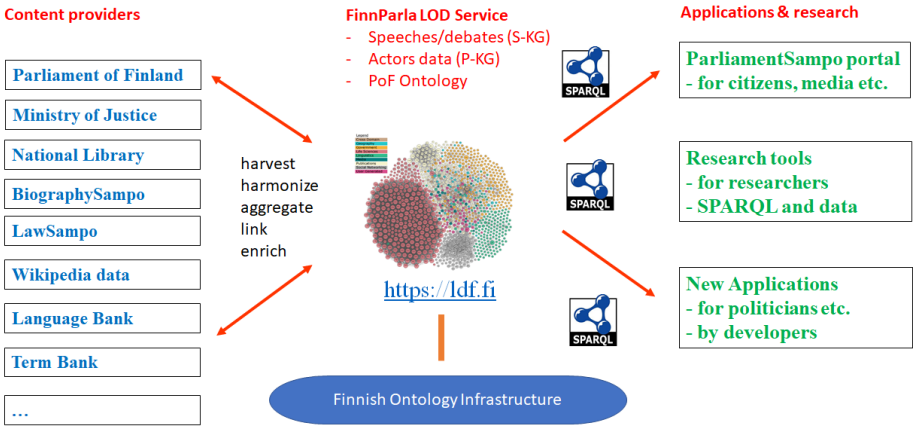


Fig. 1. Publishing model for Finnish parliamentary data in the SEMPARL project

The foundation of this work are two interlinked knowledge graphs (KG):

1. *S-KG* is a knowledge graph of all parliamentary debate speeches of Parliament of Finland (PoF) from 1907 to present time [23].
2. *P-KG* is a prosopographical knowledge graph of the Members of Parliament (MP), related other people, groups, and organizations, i.e., actors, pertaining to the parliamentary activities during the same period of time.

The two KGs are published as a LOD service called *FinnParla* about Parliament of Finland (PoF), based on an overarching ontology of PoF and the Finnish ontology infrastructure FinnONTO [10]. Fig. 1 illustrates the publishing model of SEMPARL. On the left, various content providing organizations and services are listed whose contents are transformed into, or linked with, the FinnParla LOD service in the middle. The data is used via SPARQL in research tasks and in developing applications on the right. These include the ParliamentSampo portal

under development[4]. The main data provider is PoF but also legislative data from related sources are planned to be be linked to FinnParla, such as the LawSampo data [14] from the Ministry of Justice. From the National Library, ontologies served at the Finto.fi[5] service are re-used and well as bibliographical data[6]. The Language Bank of Finland[7] contains, e.g., lots of videos of the debates, and The Helsinki Term Bank for the Arts and Sciences[8] terminological definitions pertaining to legislation and politics. The BiographySampo system contains 763 biographies of MPs as linked data as part of over 13 100 national biographies of the Finnish Literature Society. Wikipedia/Wikidata is used in various ways for enriching the FinnParla data. Possibly also media content from the Finnish Broadcasting Company Yle will be used in the project later on.

This paper introduces the prosopographical knowledge graph P-KG and addresses the following more general research question:

> *How to represent and publish prosopographical data about parliamentary actors and their activities so that the data can be used easily for Digital Humanities research?*

As an answer, the modeling principles of P-KG are presented and its transformation and publication processes are explained. It is also shown as a proof-of-concept how the LOD service can be used for Digital Humanities [21,6] research.

In the following, we first describe the original open XML data of PoF to be transformed into Linked Data. After this the RDF data model for representing parliamentary actors and their activities, as well as the transformation process are described. The produced linked data has been published as a data service using the 7-star model [11] of the Linked Data Finland platform. As a demonstration of using the data service in research, data analyses are presented. In conclusion, contributions of the work are summarized, related works are discussed, and directions for further research are outlined.

2 Parliament of Finland Actor Data

The main data source used for the P-KG is the Members of Finnish Parliament data publication available at the Parliament Open Data portal[9]. This data is regularly updated, and contains at this moment information about 2605 Members of Parliament (MP) since 1907. The person data entries are in XML format which is available in Finnish and Swedish for all the members, and in English for 202 cases.

An extract from the XML data is shown in Fig. 2. All the tags are in Finnish, and in the English version only the content is in English. A person data entry

[4] https://seco.cs.aalto.fi/projects/semparl/en/
[5] http://finto.fi/en/
[6] http://data.nationallibrary.fi
[7] https://www.kielipankki.fi/language-bank/
[8] https://tieteentermipankki.fi/wiki/Termipankki:Etusivu/en
[9] https://avoindata.eduskunta.fi/#/fi/dbsearch

```
<?xml version="1.0" ?>
<Henkilo kieliKoodi="FI" tyyppiKoodi="Kansanedustaja">
  <HenkiloNro>126</HenkiloNro>
  <EtunimetNimi>Elsi Maria</EtunimetNimi>
  <SukuNimi>Hetemäki–Olander</SukuNimi>
  <LajitteluNimi>hetemäki–olander elsi</LajitteluNimi>
  <KutsumaNimi>Elsi</KutsumaNimi>
  <MatrikkeliNimi>Hetemäki–Olander(e. Rinne, e. Hetemäki), Elsi Maria</MatrikkeliNimi>
  <Ammatti>Master of Arts, Councellor of Parliament</Ammatti>
  <SyntymaPvm>1927</SyntymaPvm>
  <SyntymaPaikka>Oulainen</SyntymaPaikka>
  ...
  < Vaalipiirit >
    < EdellisetVaalipiirit >
      < VaaliPiiri >
        <Nimi>Electoral District of Uusimaa</Nimi>
        <AlkuPvm>23.03.1970</AlkuPvm>
        <LoppuPvm>21.03.1991</LoppuPvm>
        <Tunnus>uus01</Tunnus>
      </ VaaliPiiri >
    </ EdellisetVaalipiirit >
  </ Vaalipiirit >
  ...
  <Edustajatoimet>
    <Edustajatoimi>
      <AlkuPvm>23.03.1970</AlkuPvm>
      <LoppuPvm>21.03.1991</LoppuPvm>
    </Edustajatoimi>
  </Edustajatoimet>
  ...
  <EdustajanJulkaisut>
    <EdustajanJulkaisu>
      <Nimi>Suomen vaikuttajanaisia</Nimi>
      <Vuosi>1977</Vuosi>
      <Tekijat/>
    </EdustajanJulkaisu>
  </EdustajanJulkaisut>
  ...
</Henkilo>
```

Fig. 2. Partial extract from XML data for the politician *Elsi Hetemäki-Olander*

contains biographical basic information, e.g., family name (*SukuNimi*) and given names (*EtunimetNimi*), places (*SyntymaPaikka*) and times (*SyntymaPvm*) of birth and death (if applicable), and vocations (occupations). In addition, there are detailed descriptions of the person's political, professional, and educational career. The text sample has three examples of career events: being a candidate in an electoral district (*Vaalipiiri*), being a Member of the Parliament (*Edustajatoimi*), and being a member in a parliamentary group (*Eduskuntaryhma*). These descriptions contain the label (*Nimi*) and id (*Tunnus*) of the related group and the start (*AlkuPvm*) and end (*LoppuPvm*) timestamps pertaining to the data. The data may also contain information about the publications authored by the person or about him/her. Due to privacy issues the data does not contain family-related information about the spouses and children of the politicians in contrast to many other biographical dictionaries.

3 Data Model for Parliamentary Actors and Events

To represent the biographical information about MPs and other politicians the data model presented in Fig. 3 was developed. The key idea of the model is to represent an actor's life and activities as a sequence of events (*bioc:Event*) in places (*crm:E53_Place*) and in time (*:Timespan*) with the actors (*bioc:Person*) participating in different roles (*bioc:Actor_Role*), such as *:Member*, *:Representative*, etc. The data model follows the Bio CRM [24] ontology, an extension of CIDOC CRM[10] for representing biographical information based on role-centric modeling. Bio CRM makes a distinction among attributes, relations, and events, where entities participate in different roles in a qualified manner. The namespaces used in the model are described in the figure on the left. In this extended model, there are almost 200 different roles in use. The data model has been populated by using a set of domain ontologies, such as places based on YSO places[11], groups and organizations (harvested from the data), and vocations based on the AMMO ontology [18].

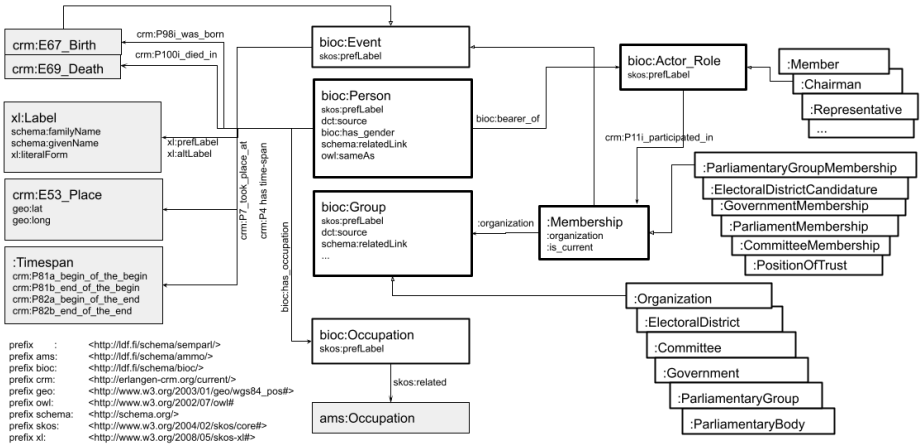


Fig. 3. Schema for the P-KG knowledge graph based on Bio CRM

For example, the XML data about the MP Elsi Hetemäki-Olander in Fig. 2 is translated into the RDF depicted in Fig. 4. Samples of extracted roles and events related to her life by the property *bioc:bearer_of* are listed in Fig. 5. As an example *event:e2044* defines her the role of being a representative relating to *event:e2043*, being a MP during the time March 23 1970 to March 21 1991.

[10] https://cidoc-crm.org
[11] https://finto.fi/yso-paikat/en/

```
PREFIX bioc:     <http://ldf.fi/schema/bioc/>
PREFIX crm:      <http://erlangen-crm.org/current/>
PREFIX event:    <http://ldf.fi/semparl/event/>
PREFIX label:    <http://ldf.fi/semparl/label/>
PREFIX occupations: <http://ldf.fi/semparl/occupations/>
PREFIX people:   <http://ldf.fi/semparl/people/>
PREFIX rdf:      <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:     <http://www.w3.org/2000/01/rdf-schema#>
PREFIX roles:    <http://ldf.fi/semparl/roles/>
PREFIX schema:   <http://schema.org/>
PREFIX semparl: <http://ldf.fi/schema/semparl/>
PREFIX skos:     <http://www.w3.org/2004/02/skos/core#>
PREFIX times:    <http://ldf.fi/semparl/times/>
PREFIX xl:       <http://www.w3.org/2008/05/skos-xl#>

people:p126 a bioc:Person ;
     crm:P98i_was_born event:b126 ;
     bioc:bearer_of event:e2044, ..., event:e2050 ;
     bioc:has_gender schema:Female ;
     bioc:has_occupation occupations:o32, occupations:o95 ;
     semparl:authored publications:b114 ;
     semparl:id "126" ;
     schema:relatedLink <https://www.eduskunta.fi/FI/kansanedustajat/Sivut/126.aspx> ;
     skos:prefLabel "Hetemäki-Olander, Elsi (1927-)"@fi ;
     xl:altLabel label:l52 , label:l53 , label:l54 ;
     xl:prefLabel label:l51 .

label:l51 a xl:Label ;
     schema:familyName "Hetemäki-Olander" ;
     schema:givenName "Elsi" ;
     skos:prefLabel "Hetemäki-Olander, Elsi"@fi .

label:l53 a xl:Label ;
     schema:familyName "Rinne" ;
     schema:givenName "Elsi Maria" ;
     skos:prefLabel "Rinne, Elsi Maria"@fi .
```

Fig. 4. Partial extract from RDF data for the politician *Elsi Hetemäki-Olander*

4 Transformation of Parliamentary Actor Data into a KG

The data contains in total 2800 person entries, i.e., instances of the class *bioc:Person*. Out of this, 2605 are MPs from the main data source Eduskunta Avoin Data[12]. This data was further enriched with data extracts from the web pages of the Finnish Government[13] and Wikidata in order to account for other people mentioned in the data and in the parliamentary speeches dataset S-KG [23] integrated with the P-KG. These ca. 200 additional resources are important people mentioned in the documents, such as Presidents of Finland, Ministers, or Parliamentary Ombudsmen[14] who have never been elected as MPs and therefore are not included in the MP database.

In addition to the people (*bioc:Person*), the groups and organizations (*bioc:Group*) mentioned in the XML data elements where extracted, disambiguated, and linked to the corresponding resources in the ontologies used. These groups contain the

[12] https://avoindata.eduskunta.fi
[13] https://valtioneuvosto.fi
[14] https://www.oikeusasiamies.fi/en/web/guest

```
event:e2044 a roles:r1 ;
    crm:P11i_participated_in event:e2043 ;
    skos:prefLabel "edustaja Hetemäki—Olander"@fi .

event:e2043 a semparl:ParliamentMembership ;
    crm:P4_has_time_span times:t814 ;
    skos:prefLabel "edustajuus 23.03.1970—21.03.1991"@fi .

event:e2050 a roles:r166 ;
    crm:P11i_participated_in event:e2049 ;
    skos:prefLabel "ehdokas Hetemäki—Olander"@fi .

event:e2049 a semparl:ElectoralDistrictCandidature ;
    crm:P4_has_time_span times:t814 ;
    semparl:is_current false ;
    semparl:organization districts:uus01 ;
    skos:prefLabel "ehdokas: Uudenmaan läänin vaalipiiri "@fi.

districts:uus01 a semparl:ElectoralDistrict ;
    skos:prefLabel "Uusimaa constituency"@en,
        "Uudenmaan läänin vaalipiiri "@fi,
        "Nylands läns valkrets "@sv .

publications:b114 a semparl:Publication ;
    crm:P4_has_time_span times:t576 ;
    skos:prefLabel "Suomen vaikuttajanaisia" .
```

Fig. 5. Samples of resourses relating to the politician *Elsi Hetemäki-Olander*

related parliamentary bodies and committees, governments, electoral districts, and furthermore also groups out of political fields, such as companies, schools, and colleges. Also references to vocations (*bioc:Occupation*) were identified and linked to the resources of the AMMO ontology of historical occupations.

As a method for knowledge extraction, patterns of regular expressions were applied to the XML data fields, especially when extracting the person name variations and expressions of time. The source data contained all terms in Finnish, in addition to a also the corresponding terms in English (1710) and Swedish (5420) were extracted, in the XML only recent data entries had translations in English. Since the main XML data came from a curated database, entities could be extracted with high precision and recall.

Table 2 summarizes the number of instances of the main classes of the data model of Fig. 3, and Table 3 lists the number of different event types extracted.

5 Prosopographical Data as a Linked Open Data Service

The prosopographical data P-KG presented above have been published on the Linked Data Finland platform[15] [11] according to the Linked Data publishing principles and other best practices of W3C [9], including, e.g., content negotiation and provision of a SPARQL[16] endpoint¨.

[15] https://ldf.fi

[16] https://www.w3.org/TR/sparql11-query/

Table 1. *

Table 2. Resources

Resource type	Count
Timespan	9168
Label	6061
Person	2801
Publication	1727
School, College	669
Place	607
Vocation	104
Parliamentary Group	89
Government	76
Committee	54
Organization	54
Electoral District	46
Parliamentary Body	38
Party	32
Ministry	12
Affiliation Group	10

Table 3. Events

Event type	Count
Career Event	14371
Position of Trust	12761
Committee Membership	6344
Municipal Position of Trust	4745
Event of Education	3712
Birth	2801
Electoral District Candidature	2205
Death	2025
Parliamentary Group Membership	1966
Government Membership	1622
Governmental Position of Trust	1621
Affiliation	1331
Parliament Membership	966
Honourable Mention	543
International Position of Trust	364
Membership Suspension	25

In our work, the "FAIR guiding principles for scientific data management and stewardship" of publishing Findable, Accessible, Interoperable, and Re-usable data are used[17]. The data can be used via the SPARQL endpoint in two ways. Firstly, the underlying SPARQL endpoint can and is being applied to custom data analyses in Digital Humanities research using tools, such as YASGUI, Google Colab, and Jupyter notebooks. Secondly, a portal called *ParliamentSampo – Finnish Parliament on the Semantic Web* is under development, a new member in the "Sampo series" of semantic portals and LOD services[18]. The portal is targeted to both researchers and the public for studying parliamentary debates, the language used, networks of Finnish politicians, and political culture. ParliamentSampo is based on the Sampo model [12] for sharing collaboratively enriched linked open data, using a shared ontology infrastructure.

The SPARQL endpoint is hosted on an Apache Jena Fuseki[19] SPARQL server. The LDF platform provides dereferencing of URIs for both human users and machines, and a generic RDF browser for technical users, which opens when a URI is visited directly with a web browser. The URI routing, content negotiation, and caching is implemented using the Varnish Cache web application accelerator[20].

[17] https://www.go-fair.org/fair-principles/
[18] https://seco.cs.aalto.fi/applications/sampo/
[19] https://jena.apache.org/documentation/fuseki2/
[20] https://varnish-cache.org

The LDF data service is based on a microservice architecture, using Docker containers[21]. Each individual component (Fuseki with the KG data and Varnish) is run in its own dedicated container, making the deployment of the services easy due to installation of software dependencies in isolated environments, enhancing the portability of the services. The data and the service are currently used internally in the SEMPARL project but will be opened by the CC BY 4.0 license to external users later on.

6 Using the SPARQL Endpoint for Data Analysis

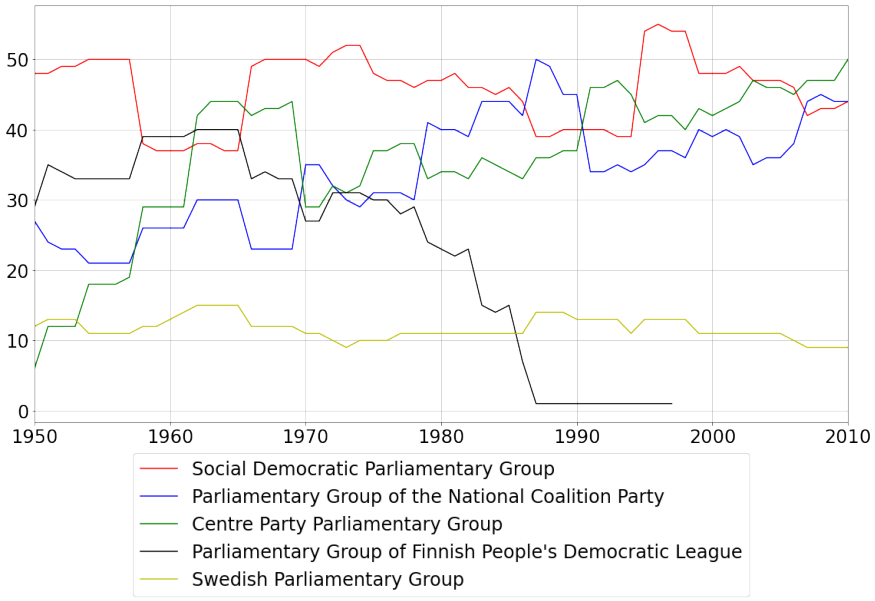


Fig. 6. Number of MPs of five most common Parliamentary Groups on a timeline

This section illustrates how the P-KG data can be used in researching the parliamentary culture in Finland, as suggested in Fig. 1.

A typical question in politics is to find out or forecast popularity of parties among the voters. Such data is available for recent times but not for historical times. By using P-KG such questions can be answered starting from 1907. For example, Fig. 6 depicts the number of MPs of the five most common Finnish parties during the years 1950–2010. The curves show how the *Social Democratic*, *National Coalition*, and *Centre Party* constantly share the top three positions.

[21] https://www.docker.com

However, the *Finnish People's Democratic League* had a significant number of representatives from 1950's to the end of 1980's; the party was later replaced by the *Left Allience*. Furthermore, during the entire period of time, the *Swedish Parliamentary Group* has had an almost constant number of MPs.

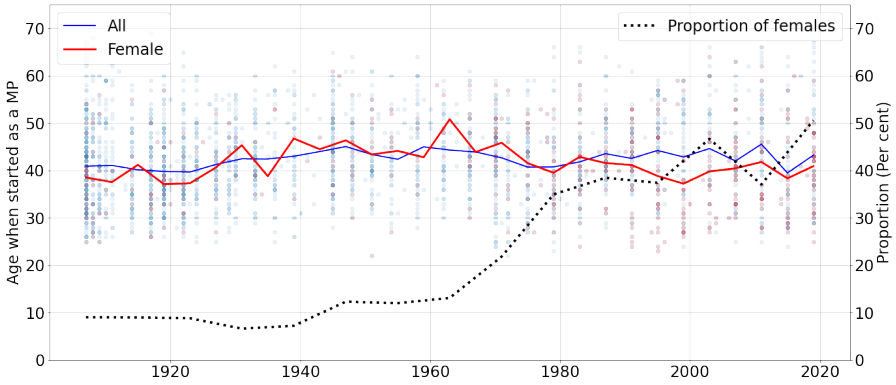


Fig. 7. Timeline with average ages of new MPs

Fig. 7 depicts on a timeline the ages of people when they were selected as MPs for the first time. The blue curve shows the average age for all MPs, and the red curve for female MPs. The values are calculated in time windows of four years. The black dotted line shows the relative proportion of female MPs in percents. It can been observed that between the years 1960 and 1980 this proportion constantly grows approximately from 10% to 40%. Generally, the average age of entering the Parliament is 42.1 which remains relatively constant during the entire timeline. However, after the 1980's the new female representatives are a few years younger than the men.

An interesting part of the P-KG is information about the vocations of the people, based on the AMMO ontology that has been aligned with the international HISCO classification[22] [25]. It provides an international comparative classification system of history of work, particularly for occupational titles in the 19th and early 20th centuries. HISCO encodes not only occupation, but also information about prestige, property, and family relations can be included. As a example of data analysis based on vocations, Fig. 8 depicts a correlation matrix between the parties and vocations of the MPs. In the figure, the rows correspond to the ten parties with most MPs and columns to their vocations. The figure shows the vocations during the entire time period from 1907 to 2021. Finland was a before the Second World War a rural country, which explains why the vocation Farmer is on the first rank. From the results it can be noticed that, e.g., *Smallholder* and *Carpenter* have been common vocations for MPs of *the Finnish People's*

[22] https://iisg.amsterdam/en/data/data-websites/history-of-work

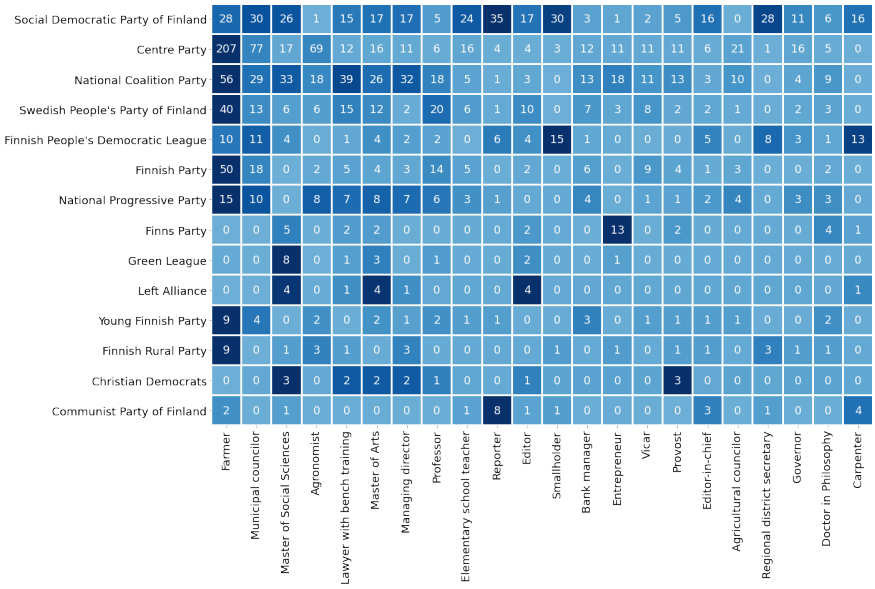


Fig. 8. Correlations between parties and vocations

Democratic League or *Provost* and *Master of Social Sciences* are common among the MPs of *the Christian Democrats*.

As a final example of data analysis, Fig. 9 depicts a correlation matrix between the parties and committees of PoF. In this figure, each row corresponds to a party and each column to a committee. The darker the cell background color is, the more members of that party haves been in the corresponding committee. Generally, the largest committee *the Grand Committee* has had a large amount of members from most of the parties. It can be noticed that, e.g., *the Finns Party* has had more members in *the Legal Affairs Committee* and *the Swedish People's Party of Finland* in *the Finance Committee*. The data model facilitates to easily generate similar visualizations of correlations between, e.g., parties, vocations, or genders.

These data analyses and visualizations were created easily by using a SPARQL query and then analysing its result with Python scripting and libraries on Google Colab Jupyter documents. According to our experiences in these and several other examples, the underlying data model and the populated data seems useful, semantically rich, and complete enough for studying political culture in versatile ways. In order to get feedback from external users, too, the data will be used in the Helsinki Digital Humanities Hackathon in May 2021 for research purposes.

Of course, the data is limited to what is openly available from PoF and to additional data and links aggregated into the P-KG from related data sources during the data transformation into RDF. When using a dataset such as P-KG,

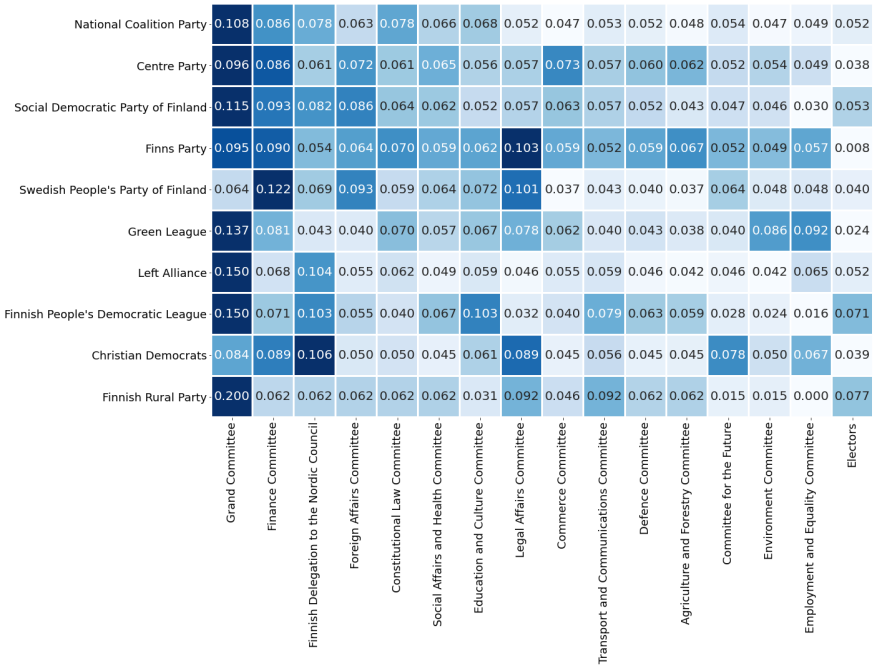


Fig. 9. Correlations between parties and committees of PoF

where much of the content has been created or transformed automatically, new kind of data literacy [19] is needed when interpreting the results. Tools based on distant reading [22] are good for finding and exploring efficiently interesting patters of information in the data but for the final interpretation and error analysis close reading is needed, too.

7 Discussion

Related Work Many national projects have transformed parliamentary data[23], such as plenary session debates [23], into structured formats and enriched the data with biographical metadata, including, e.g., the Canadian Lipad project [2] and the Norwegian Talk of Norway [20]. Linked data has also been used in some works, such as the LinkedEP about the European Parliament linked data 1999–2017 [1], the Latvian LinkedSAEIMA project [5], and the Italian Parliament[24]. Speech data can be used for analysing the language and topics of speeches (cf. e.g. [7,27,17] and also the activities of the parliament and networks of its members.

[23] See the CLARIN page www.clarin.eu/resource-families/parliamentary-corpora for a list of various national parliamentary corpora projects.

[24] http://data.camera.it/data/en/datasets/

For example, speeches of male and female MPs or other groups, such as political parties, can be analyzed and compared [4].

The P-KG is in nature a biographical dictionary even if focused on parliamentary data and events. The idea of analysing such proposographical data quantitatively, as was illustrated in section 6, have been already made for some national dictionaries of biography, such as for the British ODNB [26] and the Irish Ainm [3]. As is [16], our goal is to combine quantitative approach and distant reading methods with the qualitative approach, often based on close reading, typical to biographical research.

Contributions This paper introduced the first Linked Data model and publication of the Finnish parliament actor data, covering the whole history of PoF since 1907. In comparison to related works, the underlying data model is arguably unique in employing the semantically rich event-based ontology model presented for harmonizing data about the politicians and their lives, extending CIDOC CRM to representing prosopographical data. Our experience on developing biographical Sampo systems [13] suggests that an event-based approach is needed for integrating biographical data of different kinds instead of using only traditional document-centric models, such as Dublin Core. Furthermore, the actor data is enriched and interlinked with several additional external data sources, and is based on a national level ontology infrastructure [10] for even more extensive interlinking. The first experiments presented in using the data service for Digital Humanities research suggest that the model is fit for its purpose and can be used effectively in SPARQL queries for visualizations and parliamentary data analyses, and for creating the large Finnish parliamentary debate dataset [23] and the larger *FinnParla* LOD cloud.

Future Research Digital humanities studies are underway in the Semantic Parliament project project using the P-KG interlinked with its sister dataset S-KG about the Finnish parliamentary debates. The P-KG will also be used as part of the semantic portal *ParliamentSampo – Finnish Parliament on the Semantic Web* that is being developed based on the Sampo model [12] and the Sampo-UI framework [15].

Acknowledgements Thanks to Ari Apilo, Sari Wilenius, and Päivikki Karhula at the Parliament of Finland for co-operation, and the Semantic Parliament project team for discussions. Our work was funded by the Academy of Finland as part of the Semantic Parliament project, the EU project InTaVia: In/Tangible European Heritage[25], and is related to the COST action NexusLinguarum[26] on linguistic data science. CSC – IT Center for Science, Finland, provided computational resources for the work.

References

1. van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., Beunders, H.: The debates of the European Parliament as Linked Open Data. Semantic Web 8(2), 271–281

[25] https://intavia.eu
[26] https://nexuslinguarum.eu

(2017). https://doi.org/10.3233/SW-160227.
2. Beelen, K., Thijm, T.A., Cochrane, C., Halvemaan, K., Hirst, G., Kimmins, M., Lijbrink, S., Marx, M., Naderi, N., Rheault, L., et al.: Digitization of the canadian parliamentary debates. Canadian Journal of Political Science 50(3), 849–864 (2017). https://doi.org/10.1017/S0008423916001165.
3. Bhreathnach, Ú., Burke, C., Fhinn, J.M., Cleircín, G.Ó., Raghallaigh, B.Ó.: A quantitative analysis of biographical data from ainm, the irish-language biographical database. In: Proceedings of the Third Conference on Biographical Data in a Digital World (BD 2019) (September 2019).
4. Blaxill, L., Beelen, K.: A Feminized Language of Democracy? The Representation of Women at Westminster since 1945. Twentieth Century British History 27(3), 412–449 (2016). https://doi.org/10.1093/tcbh/hww028.
5. Bojārs, U., Darģis, R., Lavrinovičs, U., Paikens, P.: LinkedSaeima: A linked open dataset of Latvia's parliamentary debates. In: Acosta, M., Cudré-Mauroux, P., Maleshkova, M., Pellegrini, T., Sack, H., Sure-Vetter, Y. (eds.) Semantic Systems. The Power of AI and Knowledge Graphs. pp. 50–56. Springer, Cham (2019).
6. Gardiner, E., Musto, R.G.: The Digital Humanities: A Primer for Students and Scholars. Cambridge University Press, New York, NY, USA (2015), https://doi.org/10.1017/CBO9781139003865.
7. Greene, D., Cross, J.P.: Exploring the political agenda of the european parliament using a dynamic topic modeling approach. Political Analysis 25(1), 77—-94 (2017). https://doi.org/10.1017/pan.2016.7.
8. Gutierrez, C., Sequeda, J.F.: Knowledge graphs. Commun. ACM 64(3), 96—-104 (Feb 2021). https://doi.org/10.1145/3418294, https://doi.org/10.1145/3418294.
9. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool (2011), http://linkeddatabook.com/editions/1.0/.
10. Hyvönen, E., Viljanen, K., Tuominen, J., Seppälä, K.: Building a national semantic web ontology and ontology service infrastructure—the FinnONTO approach. In: Proceedings of the 5th European Semantic Web Conference (ESWC 2008). Springer (2008).
11. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) The Semantic Web: ESWC 2014 Satellite Events. ESWC 2014. pp. 226–230. Springer-Verlag (May 2014), https://doi.org/10.1007/978-3-319-11955-7_24.
12. Hyvönen, E.: "Sampo" model and semantic portals for digital humanities on the semantic web. In: DHN 2020 Digital Humanities in the Nordic Countries. Proc. of the Digital Humanities in the Nordic Countries 5th Conference. pp. 373–378. CEUR Workshop Proceedings, vol. 2612 (2020), http://ceur-ws.org/Vol-2612/poster1.pdf.
13. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., Keravuori, K.: BiographySampo - publishing and enriching biographies on the semantic web for digital humanities research. In: The Semantic Web. ESWC 2019. pp. 574–589. Springer (2019). https://doi.org/10.1007/978-3-030-21348-0_37.
14. Hyvönen, E., Tamper, M., Oksanen, A., Ikkala, E., Sarsa, S., Tuominen, J., Hietanen, A.: LawSampo: A semantic portal on a linked open data service for finnish legislation and case law. In: The Semantic Web: ESWC 2020 Satellite Events. Revised Selected Papers. pp. 110–114. Springer (2019).
15. Ikkala, E., Hyvönen, E., Rantala, H., Koho, M.: Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. Semantic Web – Interoperability, Usability, Applicability (2021), accepted.

16. Jänicke, S., Franzini, G., Cheema, M.F., Scheuermann, G.: Visual text analysis in digital humanities. Computer Graphics Forum 36(6), 226–250 (2017). https://doi.org/10.1111/cgf.12873.
17. Kettunen, K., La Mela, M.: Digging deeper into the Finnish parliamentary protocols – using a lexical semantic tagger for studying meaning change of everyman's rights (allemansrätten). In: DHN 2020 Digital Humanities in the Nordic Countries. Proc. of the Digital Humanities in the Nordic Countries 5th Conference. pp. 63–80. CEUR Workshop Proceedings, vol. 2612 (2020), http://ceur-ws.org/Vol-2612/paper5.pdf.
18. Koho, M., Gasbarra, L., Tuominen, J., Rantala, H., Jokipii, I., Hyvönen, E.: AMMO Ontology of Finnish Historical Occupations. In: Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH'19). vol. 2375, pp. 91–96. CEUR Workshop Proceedings (June 2019), http://ceur-ws.org/Vol-2375/.
19. Koltay, T.: Data literacy for researchers and data librarians. Journal of Librarianship and Information Science 49(1), 1–14 (2011). https://doi.org/DOI:10.22148/16.028.
20. Lapponi, E., Søyland, M.G., Velldal, E., Oepen, S.: The Talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016. Language Resources and Evaluation 52(3), 873–893 (Sep 2018). https://doi.org/10.1007/s10579-018-9411-5.
21. McCarty, W.: Humanities Computing. Palgrave, London (2005).
22. Moretti, F.: Distant reading. Verso Books (2013).
23. Sinikallio, L., Drobac, S., Tamper, M., Leal, R., Koho, M., Tuominen, J., La Mela, M., Hyvönen, E.: Plenary debates of the Parliament of Finland as linked open data and in Parla-CLARIN markup (March 2021), paper submitted for evaluation.
24. Tuominen, J., Hyvönen, E., Leskinen, P.: Bio CRM: A data model for representing biographical data for prosopographical research. In: Biographical Data in a Digital World (BD2017) (2017), https://doi.org/10.5281/zenodo.1040712.
25. Van Leeuwen, M.H.D., Maas, I., Miles, A.: HISCO: Historical international standard classification of occupations. Leuven University Press (2002).
26. Warren, C.: Historiography's two voices: Data infrastructure and history at scale in the oxford dictionary of national biography (ODNB). Journal of Cultural Analytics (2018). https://doi.org/DOI:10.22148/16.028.
27. Won, M., Martins, B., Raimundo, F.: Automatic extraction of relevant keyphrases for the study of issue competition. EasyChair Preprint no. 875 (EasyChair, 2019). https://doi.org/10.29007/mmk4.