Linked Data – A Paradigm Shift for Publishing and Using Biography Collections on the Semantic Web

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

[1]Semantic Computing Research Group (SeCo), Aalto University, Finland
[2]HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
[3]Finnish Literature Society
https://seco.cs.aalto.fi/projects/biografiasampo/en/

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 represent biographical data in a harmonized, semantically interoperable form, which enables 1) data enrichment by aggregating linked content from complementary, distributed, and heterogeneous data sources, as well as by reasoning, and 2) development of intelligent services using machine "understandable" data. Based on the aggregated global knowledge graph, published in a SPARQL endpoint, tooling for 1) biographical research of individual persons as well as for 2) prosopographical research on groups of people can be provided. As a demonstration of these ideas, we discuss the new in-use linked data service and semantic portal BIOGRAPHYSAMPO – *Finnish Biographies on the Semantic Web* that quickly attracted thousands of end users on the Web. This semantic portal is based on a knowledge graph extracted automatically from a collection of 13 100 textual biographies, written by 980 scholars. The texts are enriched with data linking to 16 external data sources and by harvesting external collection data from libraries, museums, and archives. Reasoning is used for query expansion and for discovering serendipitous relations between entities, such as persons and places.

1. Biographical Dictionaries on the Web

Biographical dictionaries (Keith, 2004) may contain tens of thousands of short biographies of historical persons of importance. Traditionally, such dictionaries have been published as printed book series. The Oxford Dictionary of National Biography[1] (ODNB), with more than 60 000 lives, was first published on-line in 2004, and since then major biographical dictionaries have opened their editions on the Web with search engines for finding and (close) reading biographies of interest. On-line national biographical collections include USA's American National Biography[2], Germany's Neue Deutsche Biographie[3], Biography Portal of the Netherlands[4], Dictionary of Swedish National Biography[5], and National Biography of Finland[6] (NBF).

ODNB and other early adopters of web technology started the paradigm shift in publishing and reading biographical dictionaries on the Web. We call such systems *2. generation* publications. This paper argues for taking the next step forward into *3. generation* systems, i.e., to publishing and using biographical dictionaries as Linked Data on the Semantic Web. The goal is to serve both machine and human readers, and support both close and distant reading (Shultz, June 24 2011). To demonstrate and evaluate this idea in practise, we present the new in-use system BIOGRAPHYSAMPO – *Finnish Biographies on the Semantic Web*[7] (Hyvönen et al., 2019) based on the NBF and other biography collections of the Finnish Literature Society[8]. The idea is to 1) transform textual biographies into Linked Data 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 Linked Data service and a SPARQL endpoint on the web (Heath and Bizer, 2011; Hyvönen, 2012), and to 4) create end-user applications on top of the service, including data-analytic tools and visualizations for distant reading of Big Data.

This paper considers BIOGRAPHYSAMPO from a *publishing paradigm shift* perspective, complementing our earlier papers: In (Hyvönen et al., 2019), an overview of BIOGRAPHYSAMPO from an end-user's point of view is presented; Knowledge extraction from texts is concerned in (Tamper et al., 2018); In (Tamper et al., 2019) network analysis of the biographies is in focus; In (Hyvönen and Rantala, 2019) relational search of named entities is discussed, yet another separate application perspective of the portal.

In the following, we first present the underlying "Sampo" publishing model and series of semantic portals whose new member BIOGRAPHYSAMPO is. After this the underlying knowledge graph is presented, and the new linked data based possibilities for biographical and prosopographical research are illustrated. In conclusion, related works are discussed and contributions summarized.

[1] https://www.oxforddnb.com
[2] http://www.anb.org/aboutanb.html
[3] http://www.ndb.badw-muenchen.de/ndb_aufgaben_e.htm
[4] http://www.biografischportaal.nl/en
[5] https://sok.riksarkivet.se/Sbl/Start.asp?lang=en
[6] https://kansallisbiografia.fi/english

[7] BIOGRAPHYSAMPO is available at http://biografiasampo.fi. More information and publications are available at the project homepage https://seco.cs.aalto.fi/projects/biografiasampo/en/.
[8] https://www.finlit.fi/en

2. Sampo Model for Linked Data Publishing

The ideas of the Semantic Web (SW) and Linked Data can be applied to address the problems of (semantic) data interoperability and distributed content creation at the same time, as depicted in Fig. 1. Here the publication system is illustrated by a circle. A shared semantic ontology infrastructure is situated in the middle. It includes mutually aligned metadata and shared domain ontologies, modeled using SW standards. If content providers outside of the circle provide the system with (meta)data, it is automatically linked and enriched with each other and forms a knowledge graph. For example, if metadata about a painting created by Picasso comes from an art museum, it can be enriched (linked) with, e.g., biographies from Wikipedia and other sources, photos taken of Picasso or by him, information about his wives, books in a library describing his works of art, related exhibitions open in museums, and so on. At the same time, the contents of any organization in the portal having Picasso related material get enriched by the metadata of the new artwork entered in the system. This is a win-win "business model" for everybody to join such a system; collaboration pays off.

However, the model also creates new challenges. In addition to enriching information also conflicting data from different sources may be aggregated, leading to problems of data fusion. A solution to this is to maintain provenance metadata about the primary sources (cf., e.g., (Koho et al., 2019)). This is needed also in order to promote and separate the identities of the data providers and to acknowledge their distinct contributions in the Sampo. Yet another issue is how to maintain the Sampo when aggregated data or the ontology infrastructure changes. To make this as automatic as possible, human involvement in the annotation and publishing pipeline should be minimized, as suggested in (Koho et al., 2018). However, taking the human out of the loop may lower the quality of data, and more source criticism and understanding about the limitations of the automatically annotated and aligned data is needed from the end-user (Hyvönen et al., 2019). In general, more collaboration and mutual agreements are needed between the publishers, which complicates the publishing process. Also the underlying technology needs new kind of expertise on semantic computing.

The model of Fig. 1 fits well with Linked Data idea of providing data as a service and as a live SPARQL endpoint (Heath and Bizer, 2011), on top of which independent applications can be created on the client side without server side concerns. We call this whole the Sampo[9] model (Hyvönen, 2012).

The model has been developed and tested in a series of several practical case studies, including CultureSampo[10] (2008) for cross-cultural contents, TravelSampo[11] (2011) for tourism, BookSampo[12] (2011) for fiction literature, WarSampo[13] (2015) for military history, and NameSampo[14] (2019) for toponomastic research of historical place names. Our experiences suggest that the Sampo model is a promising way to create useful systems that end-users like. For example, in 2018, BookSampo had ca. 2 million users and WarSampo 230 000.

In BIOGRAPHYSAMPO, the knowledge graph was extracted from the biography collections listed in Table 1, linked not only internally but also enriched with links to the external data sources listed in Table 2. In addition, data was harvested from 1) the art collection data of the National Gallery of Finland[15], 2) the National bibliography of Finland Fennica[16], 3) BookSampo semantic portal[17] linked data for fiction literature (Mäkelä et al., 2011), 4) the critical edition of J.V. Snellman's works (Snellman, 2002–2004)[18], and 5) the Finnish history ontology HISTO[19].

The core biographies were converted into RDF by using a natural language pipeline described in more detail in (Tamper et al., 2018). The data model used is an extension of CIDOC CRM (Doerr, 2003; Le Boeuf et al., 2019) that we call Bio CRM (Tuominen et al., 2018). In this model, the life of a person is essentially a chain of events in time and space which the person participated in different roles.

The knowledge graph was published in a Linked Data service[20] on top of which the semantic portal BIOGRAPHYSAMPO with seven application perspectives was implemented using a standard SPARQL endpoint API.

3. New Ways for Using Biographies

3.1. From Text Publishing to Tooling for DH Research

Data analysis in Digital Humanities (DH) is typically done partly by the machine, partly by the human. In visualizations, such as maps, timelines, and networks, the machine presents target data in a form from which the human user is able to make interpretations more easily. In statistics, e.g., pie charts, line charts, and histograms are used. Another type of tooling is network analysis (Newman, 2018), where different kind of connections between entities, such as family relations between persons or references between texts can be represented as graphs for visual inspection and mathematical analysis. In data-analysis and knowledge discovery, statistical or other patterns of data are searched for

[9] In Finnish mythology and the epic Kalevala, "Sampo" is a mythical artefact of indeterminate type that gives its owner richness and good fortune, an ancient metaphor of technology.

[10] https://seco.cs.aalto.fi/applications/kulttuurisampo/

[11] https://seco.cs.aalto.fi/applications/travelsampo/

[12] https://seco.cs.aalto.fi/applications/kirjasampo/

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

[14] https://seco.cs.aalto.fi/projects/nimisampo/

[15] https://www.kansallisgalleria.fi/en/avoin-data/

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

[17] http://kirjasampo.fi

[18] http://snellman.kootutteokset.fi

[19] https://seco.cs.aalto.fi/ontologies/histo/

[20] Hosted by the Linked Data Finland service http://ldf.fi.

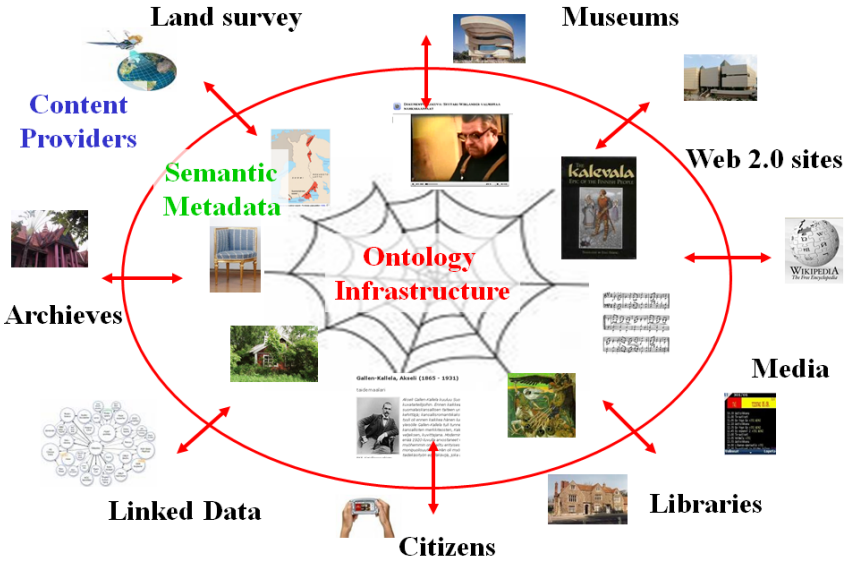

Figure 1: Sampo model for Linked Data publishing is based on a shared ontology infrastructure in the middle.

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: Core bios provided by the Finnish Literature Society.

in order to find "interesting", serendipitous (Aylett et al., 2012) new knowledge. Techniques such as topic modelling (Brett, 2012) fall in this category. The results also here typically need human interpretation, as statistical methods are usually unable to explain their results. In knowledge-based systems, knowledge structures can be used for this.

Many of the methods and tools above are well-defined and domain independent, and there are lots software packages available for using them, such as Gephi[21], R (Field et al., 2015), and various Python and JavaScript libraries. However, each of them have their own input formats and user interfaces, and need specific skills from the user. Furthermore, visualizations are crafted case by case; tools for formulating, adjusting, and comparing analysis results in some general ways would be helpful for the user.

Second generation dictionaries of biographies on the Web are used in the following traditional way: a search box or form is filled up specifying the person(s) whose biographies are searched for. Then the search button is pushed, and a list of hits is shown that can be opened for close reading by clicking. The paradigm chance of publishing biographies as linked data (third generation systems) makes it possible to build systems based on live data services, especially SPARQL endpoints. In this way, also other parties can reuse the data in their own applications. It is possible to not only publish biographies with search interfaces, but also to incorporate ready to use tooling for DH research on top of the data service. In addition, the SPARQL endpoint makes it possible to study the data by custom designed queries in situations where the ready to use interfaces are not enough for problem solving. The SPARQL API can also be used for extracting and downloading filtered subsets of data from the endpoint in different formats (e.g., CSV) to be used in external tools, such as spreadsheets, R, or Gephi. Some of these new possibilities are illustrated below by the ready to use tools of BIOGRAPHYSAMPO.

3.2. Examples: BIOGRAPHYSAMPO at Work

Problem solving in DH often has two phases, as in the prosopographical research method (Verboven et al., 2007, p. 47): First, a target group of entities in the data is selected that share desired characteristics for solving the research question at hand (in the case of prosopography, a people group is selected). Second, the target group is analyzed, and possibly compared with other groups, in order to solve the research question. Using BIOGRAPHYSAMPO based on the same pattern: First, faceted search is used for filtering out a biography or a group of them for prosopography. After this, versatile ready to use tooling can be applied for reading a single biography or for analysing groups of biographies and

[21] https://gephi.org

Data Source	# of Links	Description
Wikipedia	5806	http://fi.wikipedia.org
Wikidata	6424	http://www.wikidata.org
Fennica	4007	National Bibliography of Finland
BLF	1084	Biografiskt Lexikon för Finland
BookSampo	715	Finnish fiction literature on the Semantic Web service
WarSampo	288	Second World War LOD service and portal
ULAN	193	Union List of Artist Names Online
VIAF	2475	Virtual International Authority Files
Geni.com	5320	Family research and family tree data
Homepages	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	27586	

Table 2: External data sources linked to the BIOGRAPHYSAMPO.

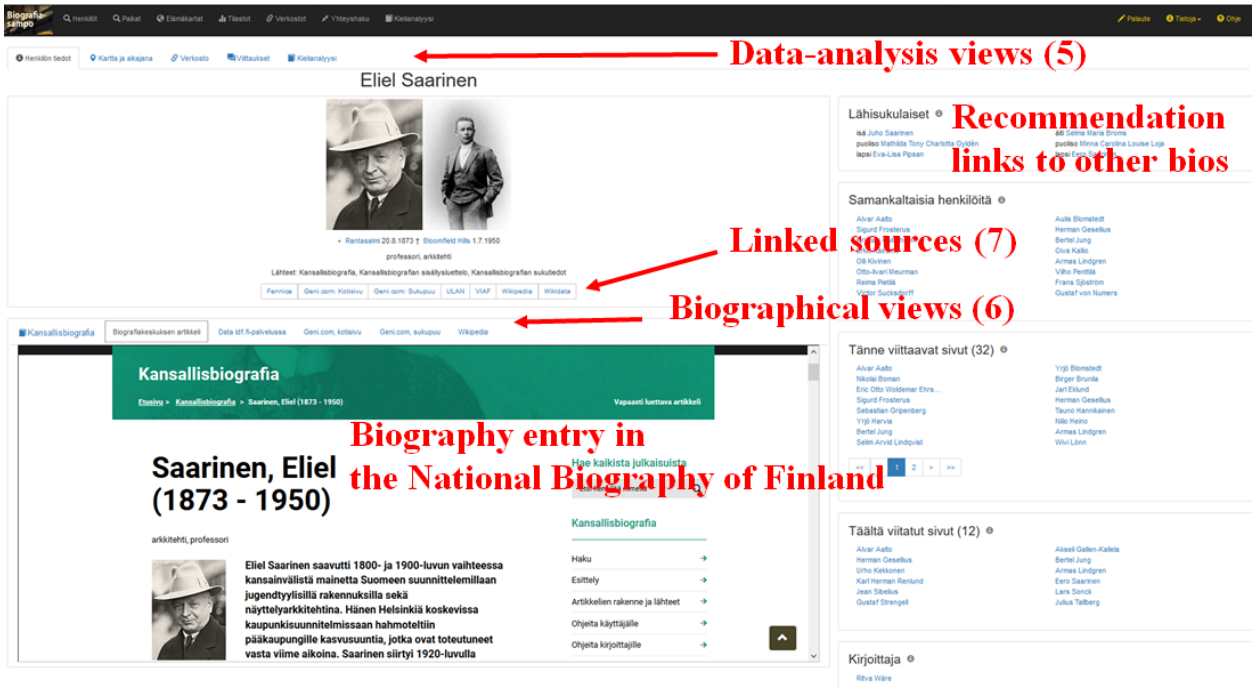

Figure 2: Homepage of Eliel Saarinen (1873–1950).

comparing them with each other or other groups.[22]

Enriching the Reading Experience After finding a biography of interest, BIOGRAPHYSAMPO provides the user with an enriched reading view of the protagonist's life by creating automatically a "homepage" for each person, based on 1) data linking and 2) reasoning. Fig. 2 shows as an example the homepage of Eliel Saarinen (1873–1950), a prominent Finnish architect. The page contains six (6) 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 homepage by the Geni.com service, and the Finnish Wikipedia article. The entry is linked to seven (7) 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. On the top of the page, there are five (5) tabs providing data-analytic views of Saarinen.

Network Analysis 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 Saarinen's life on a map and on four timelines of different event types (personal life, career, artistic or scientific creations, and accolades) for a spatiotemporal analysis.

Filtering Groups for Prosopography To support prosopography, BIOGRAPHYSAMPO employs faceted search for filtering out not only individual persons but also groups of

[22] A short video is available on the Web illustrating the ideas of BIOGRAPHYSAMPO: https://vimeo.com/328419960

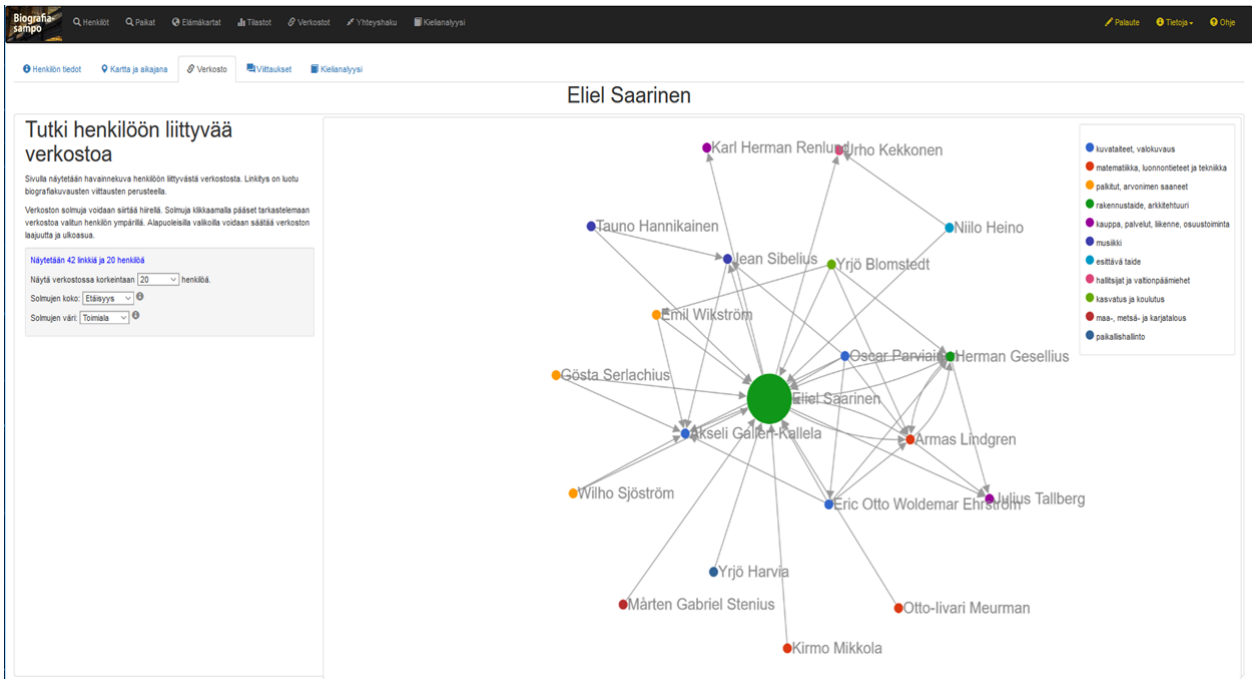

Figure 3: Egocentric network analysis of Eliel Saarinen.

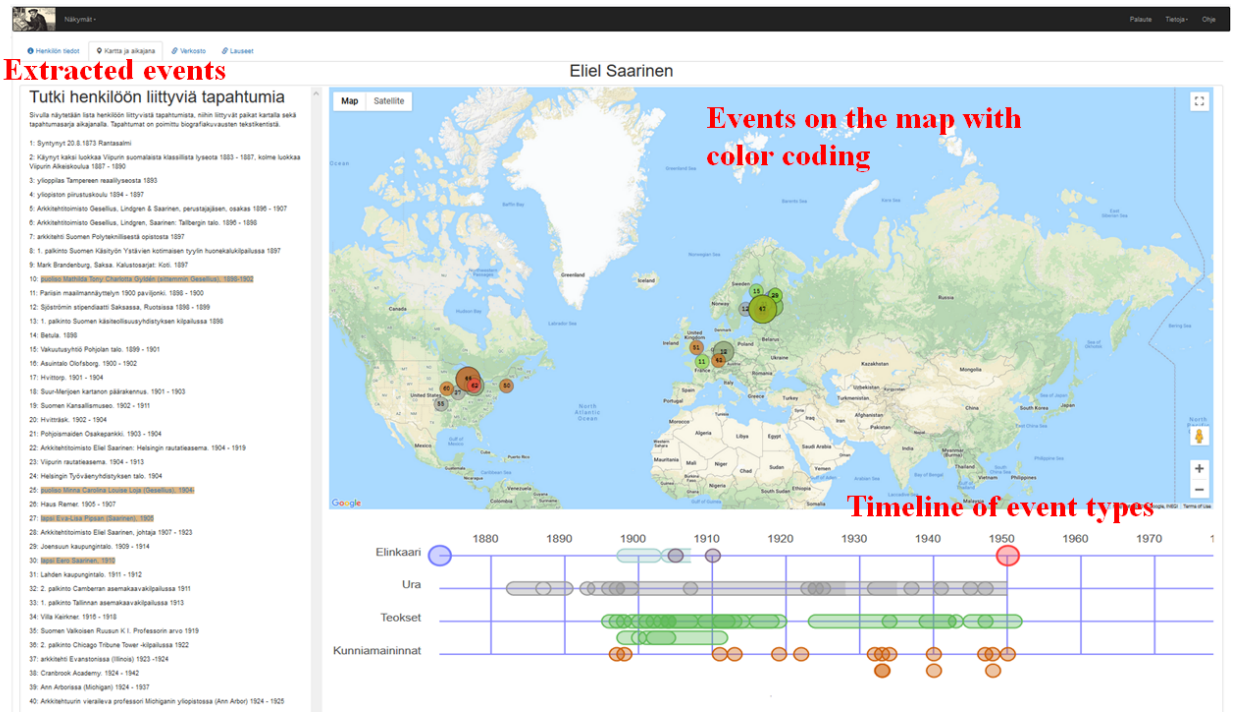

Figure 4: Spatiotempral visualization of the events in Eliel Saarinen's life.

them sharing some properties, such as profession, place of birth, place of education, working organization, etc. Once the target group has been selected, various generic data-analytic tools and visualizations can be applied to the group: 1) *Statistical tools* include histograms showing various numeric value distributions of the biographees, e.g., their ages, number of spouses and children, and pie charts visualizing proportional distributions of professions, societal domains, and working organizations. 2) *Event maps* show how different events (personal life events, career events, artistic and scientific creation events, and accolades) participated in by the biographees are distributed on maps. 3) *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).

These tools and visualization 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 generals and admirals of the Grand Duchy of Finland (1809–1917) (on the left) with the clergy (1800–1920) (on the right). With a few selections from the facets the user can see that, for some reason, quite a few officers moved the to south to die while the Lutheran ministers stayed more 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 become a chief cartographer in western Siberia where he died.

Searching for Historical Places BIOGRAPHYSAMPO also provides the user with a map search view that projects the places in which the ca. 100 000 biographical events extracted from the biographies are projected on the places where they occurred. The maps in this view are not only contemporary ones but also historical maps served by a separate historical ontology and map service Hipla.fi[23]. Many important 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 been changed, and names are now written in Russian. Using semi-transparent historical maps on top of contemporary maps solves the problem by giving a better historical context for the events.

Relational Knowledge Discovery To utilize reasoning and knowledge discovery in Linked Data, an application perspective for finding "interesting/serendipitous" (Aylett et al., 2012) connections in the biographical knowledge graph was created. This application idea is related to relational search (Lohmann et al., 2010; Tartari and Hogan, 2018). 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 rules out non-sense relations effectively and is able to create natural language explanations for the connections (Hyvönen and Rantala, 2019). 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., "Elin Danielson-Gambogi received in 1899 the Florence City Art Award". The system understands, for example, that Florence is in Italy based on the historical place ontology.

Text Analysis of Biographies The biographies can also be analyzed by using linguistic analysis, providing yet another different perspective for studying them. Both individual bios 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 the words "family" and "child", but these words are seldom used in the biographies of their male colleagues. The texts, analyzed by a natural language processing pipeline (Tamper et al., 2018), are stored in a separate knowledge graph of over 100 million triples.

[23] http://hipla.fi

4. Related Works and Contributions

Aside the business of 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 (ter Braake et al., 2015), and the proceedings of BD2017 contain more similar works (Fokkens et al., 2017b).

BIOGRAPHYSAMPO is a result of research in this area and is related to several other works. In (Larson, 2010) analytic visualizations were created based on a U.S. Legislator registry database. The work on BIOGRAPHYSAMPO is continuation to two Semantic NBF demonstrators (Hyvönen et al., 2014; Hyvönen et al., 2018), and the idea has been applied also to a historical registry of students (Hyvönen et al., 2017) and to the U.S. Legislator data (Miyakita et al., 2018). However, BIOGRAPHYSAMPO extends these systems into several new directions in terms of language technology used, the DH tooling provided, such as network analysis views, relational search, and text analysis views for studying the language of the biographies. Also more heterogeneous datasets are now studied and used.

Extracting RDF and OWL data from natural language texts has been studied in several works in semantic web research, cf., e.g., (Gangemi et al., 2017). In BiographyNet[24] (Fokkens et al., 2017a), language technology was applied to extracting entities and relations in RDF using the biographies of the Biography Portal of the Netherlands as data. This work was related to the larger NewsReader project for extracting structured data from news (Rospocher et al., 2016). 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-users, 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. Extracting and studying biographical networks has also been researched in the *Six Degrees of Francis Bacon*[25] (Warren et al., 2016) project. The statistics views and idea of analysing the biographies as a collection of texts in BIOGRAPHYSAMPO is related to (Warren, 2018) where the ODNB is analysed as an artifact. In the latter works, Linked Data is not used.

These lines of research are related to ours as they are based on the idea of extracting semantic structures from the largely unstructured biographical text collections, and on using the data for DH research in biography and prosopography. In addition and in contrast to the related works, BIOGRAPHYSAMPO employs the "Sampo model" 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

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

[25] http://www.sixdegreesoffrancisbacon.com

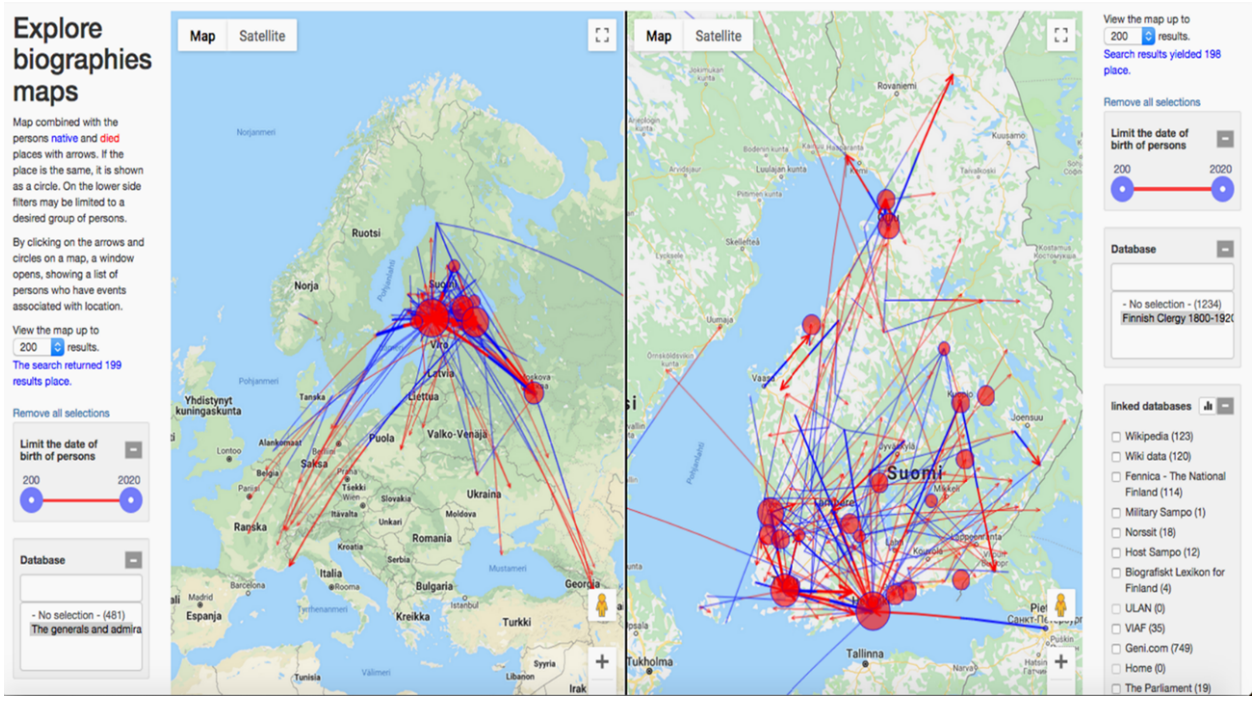

Figure 5: Comparing the life charts of two prosopographical target groups, admirals and general (left) and clergy (right) of the historical Grand Duchy of Finland (1809–1917).

critical edition, genealogical data, and various biographical data sources and semantic portals online. Another difference is that in our work, a main goal has been to develop and provide versatile DH tooling for end-users on top of a Linked Data SPARQL endpoint.

This paper presented and demonstrated the vision of a paradigm shift in publishing biography collections on the Semantic Web. The vision has also been operationalized and implemented as the semantic portal BIOGRAPHYSAMPO now in use on the Web by thousands of users. The biographical data of the portal was extracted and aggregated automatically by the computer and has not been fully validated by human experts, which would be impossible due to the amount and complexity of the big data. This is a typical situation in DH research, and calls for using more source criticism when interpreting the analyses than when dealing with human curated datasets. The quality and completeness of the BIOGRAPHYSAMPO data has not yet been analyzed formally, but our informal tests suggest that the results are very useful even if errors are also encountered. This is the price to be paid for advanced end-user services and distant reading on distributed heterogeneous biographical data.

Acknowledgements This research was part of the Severi project[26], funded mainly by Business Finland. Thanks to CSC – IT Center for Science, Finland, for computational server resources for the data service and applications.

5. References

R. S. Aylett, D. S. Bental, R. Stewart, J. Forth, and G.Wiggins. 2012. Supporting serendipitous discovery. In *Digital Futures (Third Annual Digital Economy Conference), 23-25 October, 2012, Aberdeen, UK*.

Megan R. Brett. 2012. Topic modeling: A basic introduction. *Journal of Digital Humanities*, 2(1).

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

Andy Field, Jeremy Miles, and Zoe Field. 2015. *Discovering Statistics Using R*. SAGE Publications Inc., USA.

Antske Fokkens, Serge ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, Guus Schreiber, and Victor de Boer. 2017a. BiographyNet: Extracting relations between people and events. In *Europa baut auf Biographien*, pages 193–224. New Academic Press, Wien.

Antske Fokkens, Serge ter Braake, Ronald Sluijter, Paul Arthur, and Eveline Wandl-Vogt, editors. 2017b. *BD2017 Biographical Data in a Digital World 2015*. CEUR Workshop Proceedings, Vol-1399.

Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. 2017. Semantic web machine reading with FRED. *Semantic Web Journal*, 8:873–893.

Tom Heath and Christian Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space (1st edition)*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool.

Eero Hyvönen and Heikki Rantala. 2019. Knowledge-based relation discovery in cultural heritage knowledge graphs. In *DHN2019, Digital Humanities in the Nordic Countries 2019*. CEUR Workshop Proceedings, Vol-2364.

Eero Hyvönen, Miika Alonen, Esko Ikkala, and Eetu

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

Mäkelä. 2014. Life stories as event-based linked data: Case semantic national biography. In *Proceedings of ISWC 2014 Posters & Demonstrations Track*. CEUR Workshop Proceedings, Vol-1272.

Eero Hyvönen. 2012. *Publishing and using cultural heritage linked data on the semantic web*. Morgan & Claypool, Palo Alto, CA.

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*, pages 113–119. 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.

Eero Hyvönen, Petri Leskinen, Minna Tamper, Heikki Rantala, Esko Ikkala, Jouni Tuominen, and Kirsi Keravuori. 2019. BiographySampo – publishing and enriching biographies on the semantic web for digital humanities research. In *Proceedings of the 16th Extended Semantic Web Conference (ESWC 2019)*. Springer–Verlag.

Thomas Keith. 2004. *Changing conceptions of National Biography*. Cambridge University Press.

Mikko Koho, Esko Ikkala, Erkki Heino, and Eero Hyvönen. 2018. Maintaining a linked data cloud and data service for Second World War history. In *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. 7th International Conference, EuroMed 2018, Nicosia, Cyprus*, volume 11196. Springer–Verlag.

Mikko Koho, Esko Ikkala, and Eero Hyvönen. 2019. Reassembling the lives of Finnish prisoners of the Second World War on the semantic web. In *BD-2019, Biographical Data in a Digital World 2019*. CEUR Workshop Proceedings, http://ceur-ws.org. Accepted.

Ray Larson. 2010. Bringing lives to light: Biography in context. Final Project Report, University of Berkeley, http://metadata.berkeley.edu/Biography_Final_Report.pdf.

Patrick Le Boeuf, Martin Doerr, Christian Emil Ore, and Stephen Stead, editors. 2019. *Definition of the CIDOC Conceptual Reference Model, Version 6.2.6*. ICOM/CIDOC Documentation Standards Group (CIDOC CRM Special Interest Group). http://www.cidoc-crm.org/Version/version-6.2.6.

Steffen Lohmann, Philipp Heim, Timo Stegemann, and Jürgen Ziegler. 2010. 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)*, pages 421–422. ACM.

Eetu Mäkelä, Kaisa Hypén, and Eero Hyvönen. 2011. BookSampo—lessons learned in creating a semantic portal for fiction literature. In *The Semantic Web – ISWC 201*, pages 173–188. Springer–Verlag.

Goki Miyakita, Petri Leskinen, and Eero Hyvönen. 2018. U.S. congress prosopograher - a tool for prosopographical research of legislators. In *7th International Conference, EuroMed 2018, Nicosia, Cyprus*. Springer-Verlag.

Mark Newman. 2018. *Networks*. Oxford University Press.

Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. 2016. Building event-centric knowledge graphs from news. *Web Semantics: Science, Services and Agents on the World Wide Web*, 37:132–151.

Kathryn Shultz. June, 24, 2011. What is distant reading? *New York Times*. https://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html accessed: 13 August 2018.

Johan V. Snellman. 2002–2004. *J. V. Snellman: Kootut teokset 1–24*. Ministry of Education and Culture, Helsinki, Finland.

Minna Tamper, Petri Leskinen, Kasper Apajalahti, and Eero Hyvönen. 2018. Using biographical texts as linked data for prosopographical research and applications. In *7th International Conference, EuroMed 2018, Nicosia, Cyprus, Proceedings, Part I*, pages 125–137. Springer–Verlag.

Minna Tamper, Eero Hyvönen, and Petri Leskinen. 2019. 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. Accepted.

Gonzalo Tartari and Aidan Hogan. 2018. WiSP: Weighted shortest paths for RDF graphs. In *Proceedings of VOILA 2018*. CEUR Workshop Proceedings, Vol-2187.

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 *BD-2017, Biographical Data in a Digital World 2017*, pages 59–66. CEUR Workshop Proceedings, Vol-2119.

Koenraad Verboven, Myriam Carlier, and Jan Dumolyn. 2007. A short manual to the art of prosopography. In *Prosopography approaches and applications. A handbook*, pages 35–70. Unit for Prosopographical Research (Linacre College).

Christopher N. Warren, Daniel Shore, Jessica Otis, Lawrence Wang, Mike Finegold, and Cosma Shalizi. 2016. Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks. *DHQ: Digital Humanities Quarterly*, 10(3).

Christopher N. Warren. 2018. Historiography's two voices: Data infrastructure and history at scale in the Oxford Dictionary of National Biography (ODNB). *Journal of Cultural Analytics, November 22*. doi:10.31235/osf.io/rbkdh.