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 visualizations 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 resources, 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).

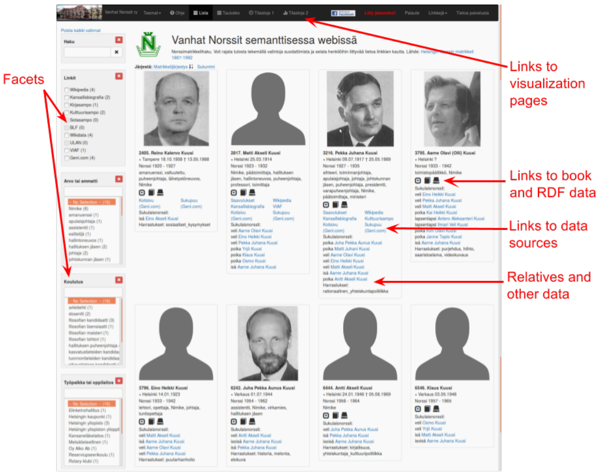

Vanhat Norssit semanttisessa webissä
Facets
Links to visualization pages
Links to book and RDF data
Links to data sources
Relatives and other data

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.

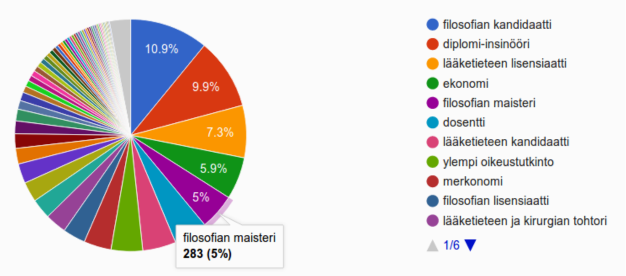


Figure 2: Pie chart showing the most common educations among high school alumni.

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

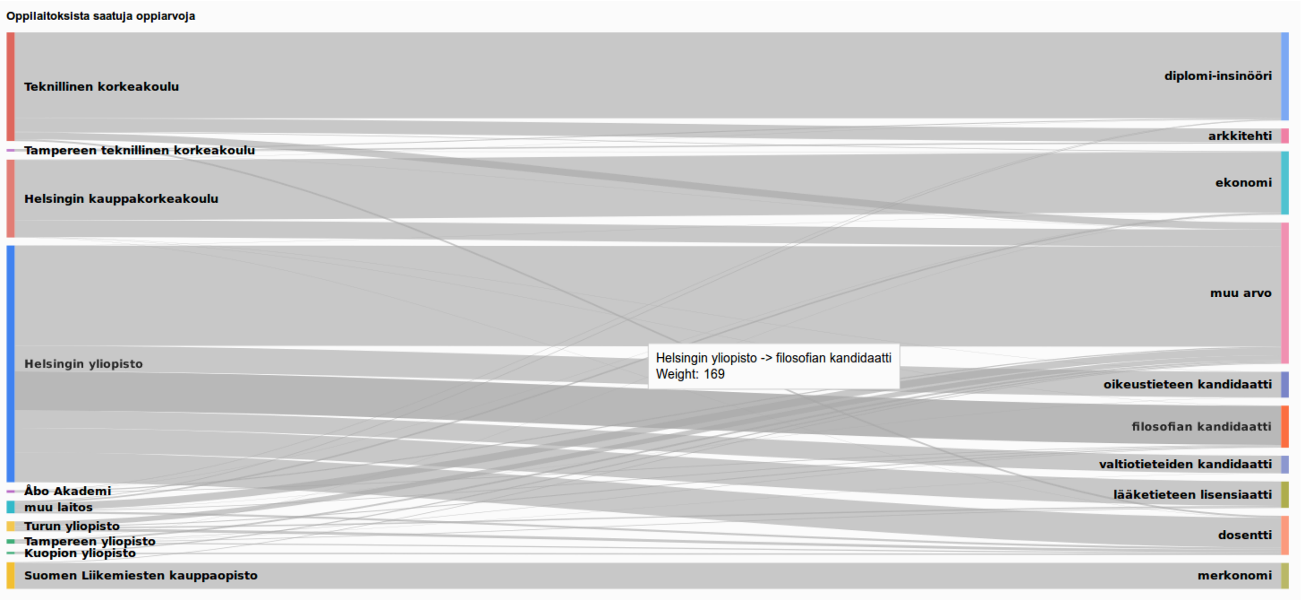


Figure 3: Sankey diagram showing the linkage between the university and the education.

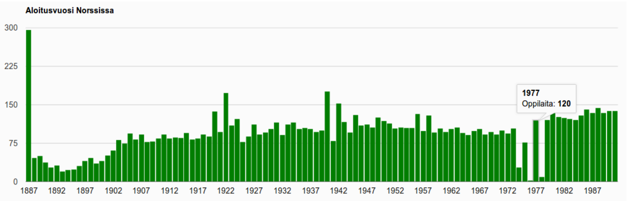


Figure 4: Column chart showing the amount of pupils by enrollment year.

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

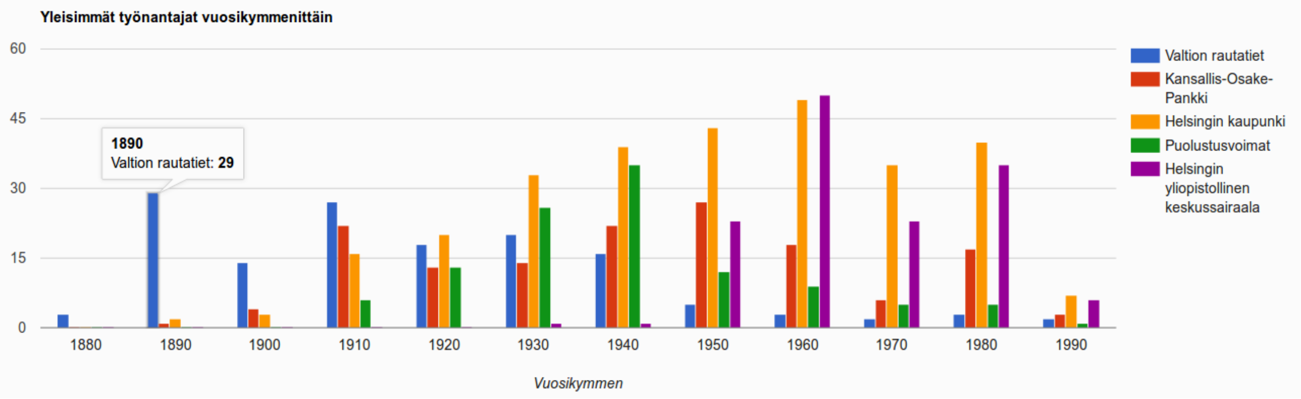


Figure 5: Column chart showing the most common employers.

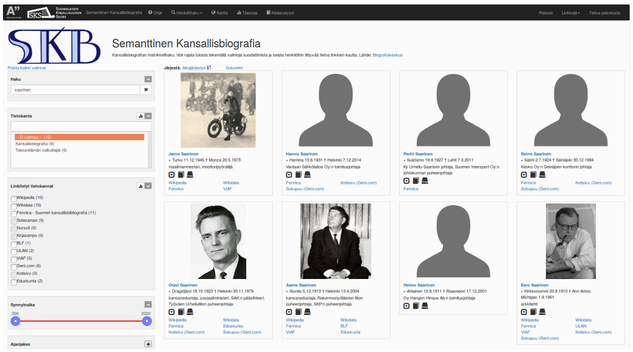
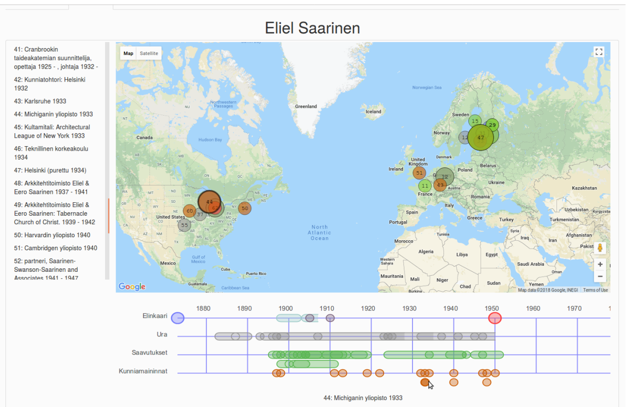


Figure 6: Main page of the Finnish National Biography.

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.


Eliel Saarinen

41: Cranbrookin taideakatemian suunnittelija, opettaja 1925 - , johtaja 1932 -

42: Kunniatohtori: Helsinki 1932

43: Karlsruhe 1933

44: Michiganin yliopisto 1933

45: Kultamitali: Architectural League of New York 1933

46: Teknillinen korkeakoulu 1934

47: Helsinki (purettu 1934)

48: Arkkitehtitoimisto Eliel & Eero Saarinen 1937 - 1941

49: Arkkitehtitoimisto Eliel & Eero Saarinen: Tabernacle Church of Christ. 1939 - 1942

50: Harvardin yliopisto 1940

51: Cambridgen yliopisto 1940

52: partneri, Saarinen-Swanson-Saarinen and Associates 1941 - 1947

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

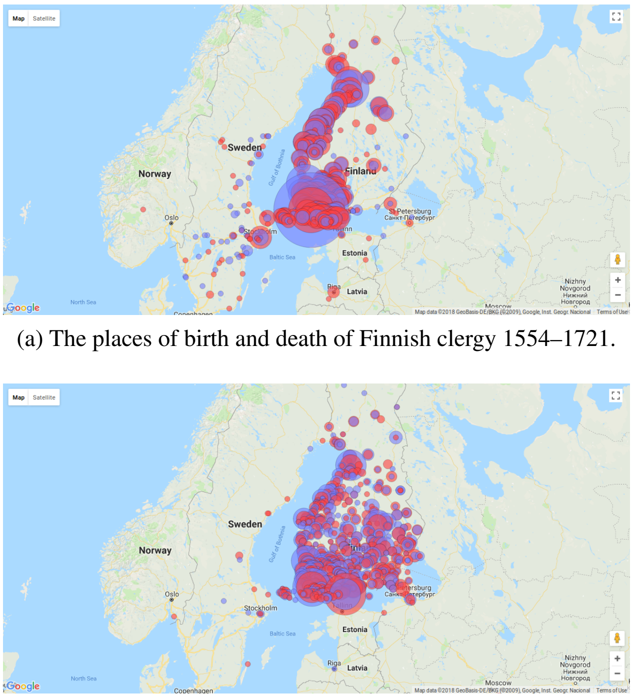
[14] http://developers.google.com/maps/

[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.

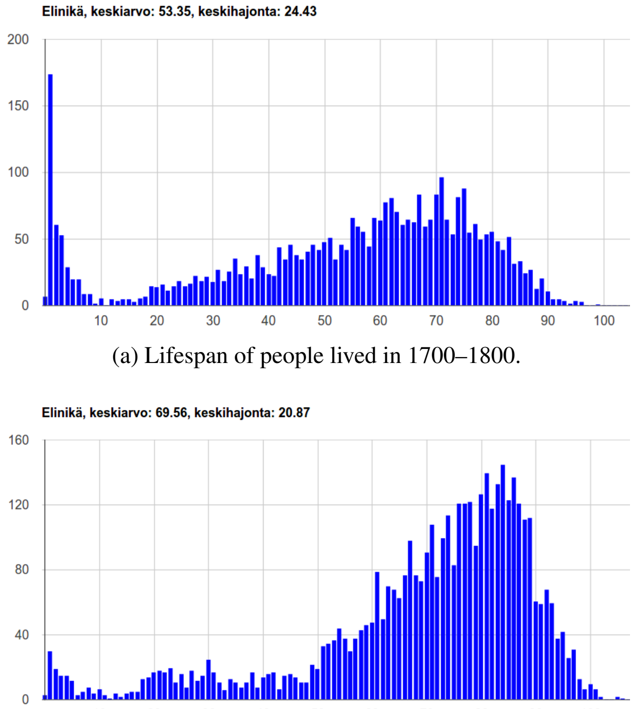


(a) The places of birth and death of Finnish clergy 1554–1721.

(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.



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

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

Figure 9: Two different views of statistical visualizations.

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 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