Hyvönen, Eero; Leskinen, Petri; Tamper, Minna; Rantala, Heikki; Ikkala, Esko; Tuominen, Jouni; Keravuori, Kirsi

BiographySampo – Publishing and enriching biographies on the semantic web for digital humanities research

Published in:
The Semantic Web - 16th International Conference, ESWC 2019, Proceedings

DOI:
10.1007/978-3-030-21348-0_37

Published: 02/06/2019

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., & Keravuori, K. (2019). BiographySampo – Publishing and enriching biographies on the semantic web for digital humanities research. In A. Zaveri, A. J. G. Gray, K. Hammar, P. Hitzler, V. Lopez, K. Janowicz, M. Fernández, & A. Haller (Eds.), *The Semantic Web - 16th International Conference, ESWC 2019, Proceedings* (pp. 574-589). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 11503 LNCS). Springer. 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.

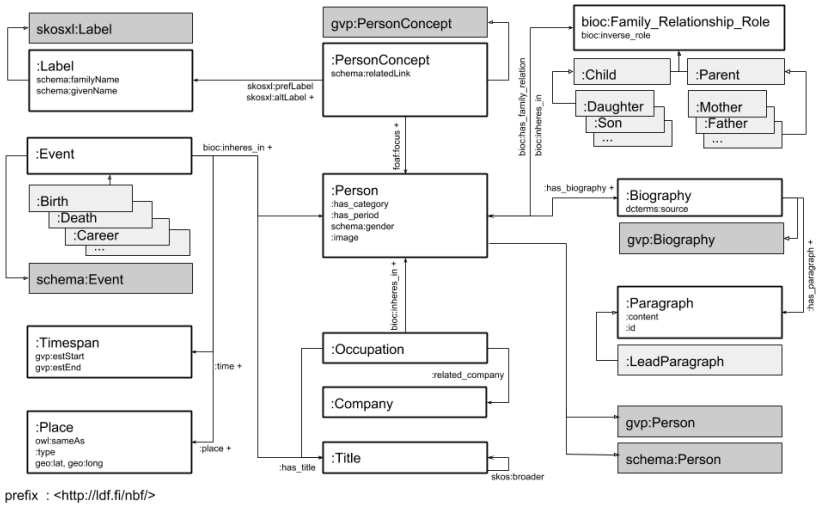

Fig. 1: Data model for BIOGRAPHYSAMPO. In addition to using *owl* and *skos(xl)* standards, namespace *bioc* refers to Bio CRM, *schema* to schema.org, *geo* to W3C Basic Geo, and *gvp* to Getty vocabulary. [36]

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.

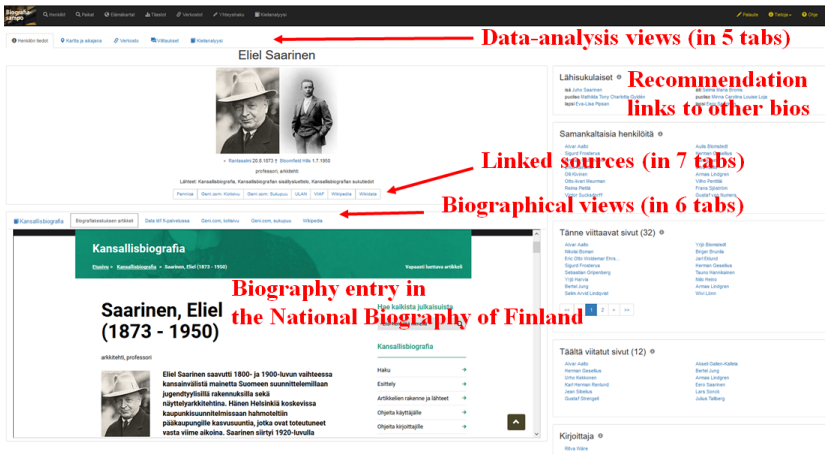

Fig. 2: Home page of Eliel Saarinen (1873–1950).

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.

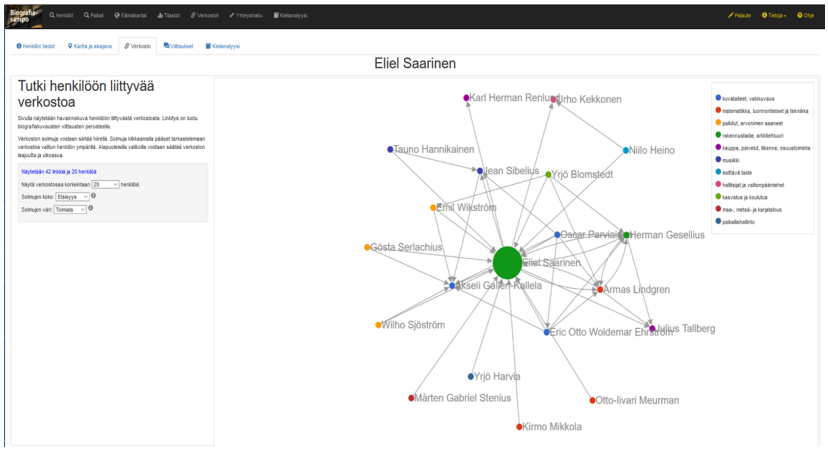

Fig. 3: Egocentric network analysis of Eliel Saarinen.

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.

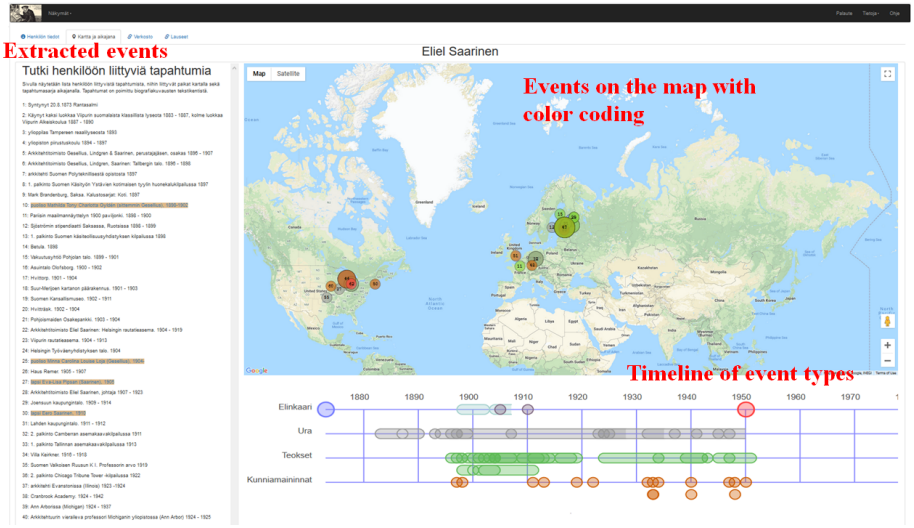

Fig. 4: Spatiotemporal visualization of the events in Eliel Saarinen's life.

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.).

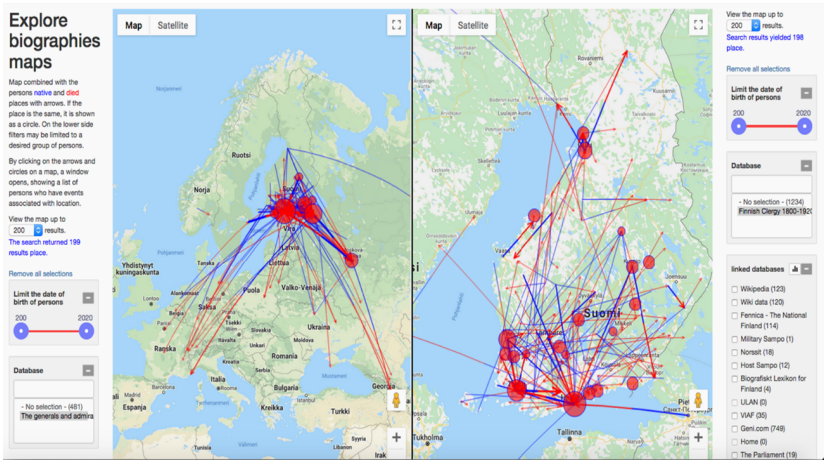

Fig. 5: Comparing the life charts of two prosopographical target groups, admirals and generals (left) and clergy (right) of the historical Grand Duchy of Finland (1809–1917).

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)