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

Eero Hyvönen¹,², Petri Leskinen¹, Minna Tamper¹, Heikki Rantala¹, Esko Ikkala¹, Jouni Tuominen¹,², and Kirsi Keravuori³

¹ Semantic Computing Research Group (SeCo), Aalto University, Finland and
² HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
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 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 2) by reasoning. 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 ‘BIOGRAFIKESKUS – 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 researchers. 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 Publishing and Using Biographical Dictionaries

Biographical dictionaries [26] may contain tens of thousands of short biographies of historical persons of importance. The Oxford Dictionary of National Biography (ODNB) [8], 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. On-line national biographical collections include USA’s American National Biography [3], Germany’s Neue Deutsche Biographie [6], France’s Nouvelle Biographie générale [7], Biography Portal of the Netherlands [4], Dictionary of Swedish National Biography [5], and National Biography of Finland⁴ [2] (NBF).

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 taking

⁴ http://biografiakeskus.fi
the next step forward, 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\(^5\) [23] based on the NBF and other biography collections of the Finnish Literature Society\(^6\). 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 [17,18], and to 4) create end-user applications on top of the service, including data-analytic tools and visualizations for distant reading [34] of Big Data.

This paper considers BIOGRAPHYSAMPO from a paradigm shift perspective, complementing four earlier publications: In [23], an overview of BIOGRAPHYSAMPO from an end-user’s point of view is presented; Knowledge extraction from texts is concerned in [36]; In [35] network analysis of the biographees is in focus; In [20] relational search of named entities is discussed, yet another separate application perspective of the portal.

In the following, we first present the underlying “Sampo” 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.

2 Publishing Cultural Heritage Linked Data: Sampo Model

---


\(^6\) https://www.finlit.fi/en
The ideas of the Semantic Web 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 metadata about CH, the data 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, 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.

Combining the infrastructure with the idea of decoupling the data services for machines from the applications for the human user creates a model for building collaborative Semantic Web applications. We call this whole the Sampo model. The model has been developed and tested in a series of several practical case studies [21], including CultureSampo (2008) for cross-cultural contents, TravelSampo (2011) for tourism, BookSampo (2011) for fiction literature, WarSampo (2015) for military history, and NameSampo (2019) for toponomastic research of historical place names.

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, 2) the National bibliography of Finland Fennica, 3) BookSampo semantic portal linked data for fiction literature [30], 4) the critical edition of J.V. Snellman’s works [1], and 5) the Finnish history ontology HISTO.

The core biographies were converted into RDF by using a natural language pipeline described in more detail in [37]. The data model used is an extension of CIDOC CRM called Bio CRM [39]. In this model, the life of a person is essentially a chain of events in which the person participated in different roles.

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

---

7 https://seco.cs.aalto.fi/applications/kulttuurisampo/
8 https://seco.cs.aalto.fi/applications/travelsampo/
9 https://seco.cs.aalto.fi/applications/kirjasampo/
11 https://seco.cs.aalto.fi/projects/nimisampo/
14 http://kirjasampo.fi
15 http://snellman.kootutteokset.fi/
16 https://seco.cs.aalto.fi/ontologies/histo/
17 Hosted by the Linked Data Finland service http://ldf.fi.
<table>
<thead>
<tr>
<th>Dataset name</th>
<th># of People</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Biography of Finland</td>
<td>6478</td>
</tr>
<tr>
<td>Business Leaders</td>
<td>2235</td>
</tr>
<tr>
<td>Finnish Generals and Admirals 1809–1917</td>
<td>481</td>
</tr>
<tr>
<td>Finnish Clergy 1554–1721</td>
<td>2716</td>
</tr>
<tr>
<td>Finnish Clergy 1800–1920</td>
<td>1234</td>
</tr>
<tr>
<td>Sum</td>
<td>13144</td>
</tr>
</tbody>
</table>

Table 1: Core bios provided by the Finnish Literature Society.

<table>
<thead>
<tr>
<th>Data Source</th>
<th># of Links</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia</td>
<td>5806</td>
<td><a href="http://fi.wikipedia.org">http://fi.wikipedia.org</a></td>
</tr>
<tr>
<td>Wikidata</td>
<td>6424</td>
<td><a href="http://www.wikidata.org">http://www.wikidata.org</a></td>
</tr>
<tr>
<td>Fennica</td>
<td>4007</td>
<td>National Bibliography of Finland</td>
</tr>
<tr>
<td>BLF</td>
<td>1084</td>
<td>Biografiskt Lexikon för Finland</td>
</tr>
<tr>
<td>BookSampo</td>
<td>715</td>
<td>Finnish fiction literature on the Semantic Web service</td>
</tr>
<tr>
<td>WarSampo</td>
<td>288</td>
<td>Second World War LOD service and portal</td>
</tr>
<tr>
<td>ULAN</td>
<td>193</td>
<td>Union List of Artist Names Online</td>
</tr>
<tr>
<td>VIAF</td>
<td>2475</td>
<td>Virtual International Authority Files</td>
</tr>
<tr>
<td>Geni.com</td>
<td>5320</td>
<td>Family research and family tree data</td>
</tr>
<tr>
<td>Homepages</td>
<td>43</td>
<td>Personal web sites</td>
</tr>
<tr>
<td>Parliament of Finland</td>
<td>631</td>
<td>Members of Parliament of Finland 1917–2018</td>
</tr>
<tr>
<td>University of Helsinki (UH) Registry</td>
<td>379</td>
<td>Students and faculty of UH in 1853–1899</td>
</tr>
<tr>
<td>Sum</td>
<td>27586</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: External data sources linked to the BIOGRAPHY SAMPO.

3 New Ways for Using Bios Based on Linked Data

3.1 From Data Publishing and Search to Tooling for DH Research

Problem solving in DH often has two phases, as in the prosopographical research method [40, 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.

The analysis in 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 statistic charts, such as pie charts, line charts, and histograms are used. Another type of tooling is network analysis [32], 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 knowl-
edge discovery statistical or other patterns of data are searched for in order to find “interesting”, serendipitous [9] new knowledge. Techniques such as topic modelling [11] 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\textsuperscript{18}, R [13], 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.

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 makes it possible to build systems that not only publish biographies with search interfaces, but also incorporate ready to use tooling for DH research on top of them. In addition, the idea of publishing the data as a service in a 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. In addition, other parties can reuse the data in novel ways in their own application—\textsc{BiographySampo} itself demonstrates this by reusing several available SPARQL endpoints on the web (cf. Table 2). In the following, some of these new possibilities are illustrated by considering some functionalities of \textsc{BiographySampo} as examples.

3.2 \textbf{Examples: Biographical at Work}

\textsc{BiographySampo} is used by the following basic pattern: 1) faceted search (including the traditional text search as a facet) is used for filtering our a biography or a group of them for prosopography. 2) Versatile ready to use tooling can be applied for reading a single biography or for analysing groups of biographies and comparing them with each other.\textsuperscript{19}

\textbf{Enriching the Reading Experience} After finding a biography of interest, \textsc{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 Elie 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 by based on their linguistic content.

\textsuperscript{18} https://gephi.org/

\textsuperscript{19} A short video is available on the Web illustrating the ideas of \textsc{BiographySampo}: https://vimeo.com/328419960.
Fig. 2: Homepage of Eliel Saarinen (1873–1950).

Fig. 3: Egocentric network analysis of Eliel Saarinen.

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 for events of different types (personal life, career, artistic or scientific creations, and accolades) for a spatiotemporal analysis.

**Filtering Groups for Prosopography** To support prosopography, BIOGRAPHY-SAMPO employs faceted search for filtering out not only individual persons but also groups of 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 the he was promoted to 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. 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 digitized 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" connections in the biographical knowledge graph was created. This application idea is related to relational search [29,38]. 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 [25]. 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" and "Robert Ekman created in 1844 the painting 'Landscape in

Fig. 5: Comparing the life charts of two propographical target groups, admirals and general (left) and clergy (right) of the historical Grand Duchy of Finland (1809–1917).
“Subiaco’ depicting a place in Italy”. Based on the underlying historical place ontology, the system understands, for example, that Florence is in Italy.

**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 male members. The texts, analyzed by a natural language processing pipeline [37], are stored in a separate knowledge graph of over 100 million triples.

### 4 Discussion: 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 [10], and the proceedings of BD2017 contain more similar works [15].

**BiographySampo** is a result of research in this area and is related to several other works. In [27] analytic visualizations were created based on U.S. Legislator registry data. The work on BiographySampo was influenced by the early Semantic NBF demonstrator [19] and its follow-up prototype [24], whose software has been applied also to a historical registry of students [22] and to the U.S. Legislator data [31]. However, BiographySampo extends these systems into several new directions in terms of 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 used.

Extracting RDF and OWL data from natural language texts has been studied in several works in semantic web research, cf. e.g. [16]. In BiographyNet project[21] [14] language technology was applied for 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 [33]. Extracting and studying biographical networks has also been researched in the Six Degrees of Francis Bacon[22] [41] project. This line of research is similar to ours, based on the idea of extracting semantic 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 proving 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" [21], where the data is enriched through a shared content infrastructure by related external heterogeneous

---

21 http://www.biographynet.nl/
22 http://www.sixdegreesoffrancisbacon.com/
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.

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 BIOGRAPHY/AMPO 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 data has not yet been 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 more advanced end-user services and distant reading.

Acknowledgements

This research was part of the Severi project\textsuperscript{23}, funded mainly by Business Finland. Thanks to CSC – IT Center for Science, Finland, for providing computational server resources for the application.

References

7. Nouvelle Biographie générale (2017), https://fr.wikipedia.org/wiki/Nouvelle_Biographie_g%C3%A9n%C3%A9rale

\textsuperscript{23} http://seco.cs.aalto.fi/projects/severi