BIOGRAPHYSAMPO – A Paradigm Shift 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. Firstly, a biographical dictionary on the web should provide the end user with an enhanced reading experience of biographies by enriching them with data linking and reasoning. Secondly, the web publication should include not only biographies for humans to read but also versatile tooling for 1) biographical research of individual persons as well as for 2) prosopographical research on groups of people. To support these arguments, we present the designing principles and the implementation of the semantic portal "BiographySampo – Finnish Life Stories on the Semantic Web" especially from the end user’s point of view. The system is based on a Linked Data service and knowledge graph extracted automatically from a collection of 13 100 textual biographies, written by 900 researchers. The texts are enriched with data linking to 16 external data sources and by harvesting external collection data from libraries, museums, and archives. The portal, consisting of seven different interlinked application perspectives, was released on September 27, 2018, for free public use⁴ for Digital Humanities researchers and the general public.

1 Publishing and Using Biographical Dictionaries

Biographical dictionaries [20] may contain tens of thousands of short biographies of historical persons of importance. The Oxford Dictionary of National Biography (ODNB) [6], 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 [1], Germany’s Neue Deutsche Biographie [4], France’s Nouvelle Biographie générale [5], Biography Portal of the Netherlands [2], Dictionary of Swedish National Biography [3], and National Biography of Finland⁵ [21] (NBF). While the biographical dictionaries had an

⁴ http://biografiasampo.fi
⁵ http://biografiakeskus.fi
unconcealed nationalist agenda well into the 20th century, the contemporary on-line national biographies have all made an effort to include groups previously ignored by national history. Pioneering women in all fields are included, as well as many marginal and minority groups.

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 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 Life Stories on the Semantic Web” 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 [13,14], and to 4) create end-user applications on top of the service, including data-analytic tools and visualizations for distant reading [29] of Big Data.

In this paper, BIOGRAPHYSAMPO is considered from the DH end user’s perspective. We first discuss the underlying knowledge graph, and then illustrate possibilities for biographical and prosopographical research that the new approach facilitates. In conclusion, related works are discussed and contributions summarized.

2 Creating the Knowledge Graph and Linked Data Service

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\(^7\), 2) the National bibliography of Finland Fennica\(^8\), 3) BookSampo semantic portal\(^9\) linked data for fiction literature [25], 4) the critical edition of J.V. Snellman’s works [28]\(^10\), and 5) the Finnish history ontology HISTO\(^11\).

The core bios were converted into RDF by using a natural language pipeline described in more detail in [30]. The data model used is an extension of CIDOC CRM [9,23] called Bio CRM [32]. 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\(^12\) on top of which the semantic portal BIOGRAPHYSAMPO with seven application perspectives was implemented using a standard SPARQL endpoint API.

---

\(^6\) https://www.finlit.fi/en
\(^7\) https://www.kansallisgalleria.fi/en/avoin-data/
\(^9\) http://kirjasampo.fi
\(^10\) http://snellman.kootuttekset.fi/
\(^11\) https://seco.cs.aalto.fi/ontologies/histo/
\(^12\) 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><strong>Sum</strong></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>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><strong>Sum</strong></td>
<td>27586</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: External data sources linked to the BIOGRAPHYSAMPO.

### 3 New Ways for Using Bios

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. BIOGRAPHYSAMPO challenges this approach in the following ways:

#### 3.1 Enriching the Reading Experience

After finding a biography of interest, BIOGRAPHYSAMPO provides the user with an enriched reading view of the protagonist’s life by creating automatically a "homepage" for each person, based on 1) data linking and 2) reasoning. Fig. 1 shows as an example the homepage of Eliel Saarinen (1873–1950), a prominent Finnish architect. The page contains six (6) tabs providing different biographical views of the person, here...
two pages based on the NBF, data at the Linked Data Finland service, a genealogical family tree and homepage by the Geni.com service, and the Finnish Wikipedia article. The entry is linked to seven (7) external data sources on the web. On the right, recommendation links to related biographies are given, e.g., to similar biographies by based on their linguistic content.

Fig. 1: Homepage of Eliel Saarinen (1873–1950).

Fig. 2: Egocentric network analysis of Eliel Saarinen.
On the top of the page, there are five (5) tabs providing data-analytic views of Saarinen. For example, Fig. 2 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. 3, another tab visualizes the international events of Saarinen’s life on a map and a timeline for a spatiotemporal analysis.

### 3.2 Filtering Groups for Prosopography

The prosopographical research method [33, 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. To support prosopography, BIOGRAPHYSAMPO employs faceted search for filtering out not only individual biographies/persons but also groups of bios/persons 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 protagonists, 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 (births, deaths, career events, artistic creation events, and accolades) participated by the protagonists 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).
Fig. 4: 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).

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. 4 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 soldiers moved to the south to die while the priests 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.

3.3 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.

http://hipla.fi
3.4 Relational Knowledge Discovery

To utilize reasoning and knowledge discovery in Linked Data, an application perspective for finding "interesting/serendipitous" [7] connections in the biographical knowledge graph was created. This application idea is related to relational search [24,31]. 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 [19]. 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 Subiaco' depicting a place in Italy".

3.5 Text Analysis of Biographies

The biographies can 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 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 [30], 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 [8], and the proceedings of BD2017 contain more similar works [11].

BiographySampo is a result of research in this area and is related to several other works. In [22] analytic visualizations were created based on U.S. Legislator registry data. The work on BiographySampo was influenced by the early Semantic NBF demonstrator [15] and its follow-up prototype [18], whose software has been applied also to a historical registry of students [17] and to the U.S. Legislator data [26]. 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. [12]. In [10] language technology was applied for extracting entities and relations in RDF using Dutch biographies as data in BiographyNet. This work was part of the larger NewsReader project extracting structured data from news [27]. 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 tools for biography and prosopography. In addition and in contrast to the related works, BiographySampo employs the "Sampo-model" [16], 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.

This paper presented and demonstrated the vision of a paradigm shift in publishing biography collections on the Semantic Web. The vision has also been operationalized and implemented as the semantic portal BIOGRAPHYSAMPO now in use on the Web. 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[^14], funded mainly by Business Finland.

**References**

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


