Creating and Using Biographical Dictionaries for Digital Humanities Based on Linked Data: A Survey of Web Services in Use in Finland

Eero Hyvönen[1,2]

[1] *Semantic Computing Research Group (SeCo), Aalto University, Finland, https://seco.cs.aalto.fi*
[2] *Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland, https://heldig.fi*

Abstract
This paper overviews work on creating Linked Open Data (LOD) ontology and data services and applications for publishing and using biographical collections on the Semantic Web, both for Digital Humanities research and for the general public. The focus is on presenting a series of six LOD services and related biographical portals in use in Finland, based on biographies and person registries of historical people. These web services were developed based on a vision of four technological generations of publishing biographies. The lessons learned in the presented work contributed to the so-called Sampo model for publishing and using Cultural Heritage data in a collaborative way using a shared ontology infrastructure. The six biographical Sampo systems discussed have had up to a million of users on the Web suggesting feasibility of the proposed model and tools used in creating them.

Keywords
Semantic Web, Biographies, Prosopography, Digital Humanities, Linked Open Data

1. Biographies as Linked Data on the Semantic Web

Biographical data are typically strongly interlinked, but published in heterogeneous, distributed local data silos, making it difficult to utilize the data on a global level. Both tangible and intangible Cultural Heritage (CH) data is involved. Furthermore, the content is usually available only for humans to read, and not as data for Digital Humanities (DH) analyses and application development.

This paper argues that these challenges can be addressed effectively using the *Sampo model* [1] and related user interface (UI) tools, such as the SPARQL Faceter [2] and Sampo-UI [3] framework in our case. The paper surveys a line of research where this approach was applied to developing a series of six biographical/prosopographical applications available on the Semantic Web in Finland: WarSampo[1] [4], Norssi Alumni[2] [5], U.S. Congress Proposographer[3] [6], Biogra-

Biographical Data in a Digital World (BD 2022), July 25th, 2022, virtual, Tokyo, Japan
eero.hyvonen@aalto.fi (E. Hyvönen)
https://seco.cs.aalto.fi/u/eahyvone/ (E. Hyvönen)
0000-0003-1695-5840 (E. Hyvönen)

https://doi.org/10.3986/9789610508120_04

[1] Available at: https://sotasampo.fi; Project home: https://seco.cs.aalto.fi/projects/sotasampo/en/
[2] Portal: https://norssit.fi/semweb; Project home: https://seco.cs.aalto.fi/projects/norssit/
[3] Portal: https://semanticcomputing.github.io/congress-legislators/

phySampo[4] [7, 8], WarVictimSampo 1914–1922[5] [9], and AcademySampo[6] [10, 11]. Table 1 presents a summary of these systems online with their year of publication, application domain, number of users, size of the knowledge graph, and a list of primary data owners.

Table 1
Six biographical Sampo portals and LOD services for Digital Humanities; distinct user counts (site visits) by Google Analytics in 2021 October

Portal	Year	Domain	# Users	# Triples	Primary data owners
WarSampo	2015–2019	World War II	1 000 000	14M	National Archives, Defense Forces, and others, Finland
Norssi Alumni	2017	Person registry	unknown	0.47M	Norssi High School alumni organization Vanhat Norssit
U.S. Congress Prosopographer	2018	Politicians	unknown	0.83M	U. S. Congress Legislator data
BiographySampo	2019	Biographies	50 000	5.56M	Finnish Literature Society
WarVictimSampo 1914–1922	2019	Military history	29 000	9.96M	National Archives of Finland
AcademySampo	2021	Finnish Academics	8200	6.55M	University of Helsinki and National Archives, Finland

This paper overviews these systems by relating them to generations of publishing biographies and by explaining their underlying design principles. In the following, related works are first discussed (Section 2) and a vision of four consecutive generations of publishing biographical information is presented. The work on the biographical Sampo systems has been driven by this vision of developing more and more useful and intelligent biographical publishing systems for both researchers and the general public. The lessons learned have gradually evolved into the so-called Sampo model, a set of six principles for developing LOD services and semantic portals on top of them, discussed in Section 4. The paper ends with a summary on contributions and challenges of the presented work.

2. Related Work

Biographies are an important source of information for researchers across various disciplines with an interest in the history. [12] Biographical dictionaries are scholarly resources used not only by the academic community but also the by the public. Such dictionaries typically start with narrative text describing the life of a biographee followed by a structured synopsis of his/her basic biographical facts, such as family relations, education, works, career events, and so on.

In addition textual dictionaries there are also prosopographical databases or data services

[4]Portal: https://biografiasampo.fi; Project home: https://seco.cs.aalto.fi/projects/biografiasampo/en/
[5]Portal: https://arkisto.sotasurmat.fi ; Project home: https://seco.cs.aalto.fi/projects/sotasurmat-1914-1922/en/
[6]Portal: https://akatemiasampo.fi; Project home: https://seco.cs.aalto.fi/projects/yo-matrikkelit/

available on the Web. For example, in Austria there is the dictionary OEBL[7] online serving biographical texts and also the Austrian Prosopographical Information System APIS[8] as a further advancement with structured, linked data. From databases and data services online, structured data be exported from the service and/or reused via application programming interfaces (API).

An example of a biographical dictionary is the Oxford Dictionary of National Biography (ODNB)[9] with more than 60 000 lives. It was published in print and online in 2004. Today many dictionaries are available on the Web. These include USA's American National Biography[10], Germany's Neue Deutsche Biographie[11], Biography Portal of the Netherlands[12], The Dictionary of Swedish National Biography[13], and the National Biography of Finland[14] (NBF). There are also many "who is who" services online, and Wikipedia contains lots of short biographies with lots of data available in DBpedia[15] and Wikidata[16].

Biographical collections can be used to study the underlying historical world. However, the texts, the language used, and the biographical collection as a whole can also be studied from a different, historiographical perspective as an artifact reflecting its own time, the editorial values and biases in selecting the biographees, the authors' perspectives, and also from linguistic points of view. Such analyses have been already made for some national dictionaries of biography, e.g., for the British ODNB [14], the Irish Ainm [15], Biography Portal of the Netherlands/BiographyNet [16], APIS in Austria [17], and the National Biography of Finland/BiographySampo [7, 8]. There are also related studies using, e.g., Wikipedia articles as the data source [18, 19].

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 [20], and the proceedings of BD2017 contain more similar works [21]. In [22], analytic visualizations were created based on U.S. Legislator registry data. The idea of biographical network analysis was developed in the Six Degrees of Francis Bacon system[17] [23, 24] that utilizes data of the Oxford Dictionary of National Biography. Network analyses and visualizations based on biographies have been presented also in [25, 10].

Extracting Linked Data [26] from texts [27] has been studied in several works, cf., e.g., [28, 29]. In [16] language technology was applied for extracting entities and relations in RDF using Dutch biographies in the BiographyNet[18]. This work was part of the larger NewsReader project[19] extracting data from news [30]. The problem of extracting (linked) data from biographical texts has also been studied when transforming the biographies of AcademySampo [31] and

[7] https://www.biographien.ac.at/
[8] https://apis.acdh.oeaw.ac.at/
[9] http://global.oup.com/oxforddnb/info/
[10] http://www.anb.org/aboutanb.html
[11] http://www.ndb.badw-muenchen.de/ndb_aufgaben_e.htm
[12] http://www.biografischportaal.nl/en
[13] https://sok.riksarkivet.se/Sbl/Start.aspx?lang=en
[14] http://kansallisbiografia.fi [13]
[15] https://www.dbpedia.org/
[16] https://www.wikidata.org/
[17] http://www.sixdegreesoffrancisbacon.com
[18] http://www.biographynet.nl/
[19] http://www.newsreader-project.eu/

BiographySampo [32] into LOD. BiographyNet focuses more on the challenges of natural language processing and managing the provenance information of data from multiple sources, while our focus in Sampo systems is on providing the end user with intelligent search and browsing facilities, enriched reading experience, and easy to use data-analytic tooling for biography and prosopography. The Austrian Prosopographical Information System (APIS) [17, 33, 34] is a virtual research environment that transforms text collections to machine readable formats and enables the use of natural language processing based methods to enrich the documents by extracting and linking information in them. The system has been used to transform and to study the collection of Austrian Biographical Dictionary 1815–1950 (ÖBL). Similarly to the Sampo systems, the APIS can be used to analyze and visualize datasets using for example network analysis methods.

3. Generations of Publishing Biographies

Table 2
Generations of Publishing Biographies

1. Generation	Engravings and printed texts
2. Generation	Biographies online for close reading
3. Generation	Biographies as data for data analysis and distant reading
4. Generation	Automatic knowledge discovery and AI

The idea of publishing biographies has evolved in generations [35] (cf. Table 2). First, life stories were published as texts engraved, e.g., in tomb stones in China and in rune stones in Scandinavia, and later as hand-written or printed texts (1. generation). Publishing biographies on the Web for close reading can be seen as the next 2. generation in the 90's. These systems can be referred to as dictionaries of biography on the Web. Here the biographies are provided for humans to read independently from place and time. Search engines are used for finding persons and texts of interest, and by browsing hypertext links additional recommended sources of information can be found. However, in 2. generation systems the data can be read only by the human user, and not by machines that only communicate the contents: the underlying data is not provided for computational analyses and application development. As a remedy, the data underlying the biographies can be published as a structured prosopographical database to be used in applications via APIs, e.g., for 2. generation dictionaries. In BiographySampo [7] the idea of publishing and using biographies as linked data was argued as a new paradigm change. Linked data can not only be used for data search and exploration as in 2. generation systems but also for data-analytic Digital Humanities (DH) research [36]. (Linked) data-based biographical publications can be seen as 3. generation systems. Arguably knowledge discovery, based on Artificial Intelligence, could be the next step ahead to the 4. generation of publishing biographies. Here the computer by itself is able to find new research questions in the data, solve them, and even explain the solutions to the human. First steps towards 4. generation systems in the case of BiographySampo are discussed in [35, 37].

Our work on biographical linked data started 2013 by designing a demonstrator [38] based on

the short biographies of the National Biography of Finland (NBF)[20]. The research hypotheses of this system was that "the reading experience can be enhanced by enriching the biographies with additional life time events, by proving the user with a spatio-temporal context for reading, and by linking the text to additional contents in related datasets". The demonstrator was a 2. generation system although some map-based visualizations were created. The lives of the biographees were modeled as sequences of spatio-temporal events using linked data and CIDOC CRM[21] [39], an approach that has been used constantly in the later biographical Sampo systems.

The NBF demonstrator and its approach of using an event-based model for biographical linked data lead us to develop the system *WarSampo – Finnish World War II on the Semantic Web*[22] (online since 2015 with several new application perspectives published in 2016–2019) [4]. A key idea in WarSampo is to reassemble the life stories of the WW2 soldiers using data linking from different data sources. Biographical/prosopograohical data was represented using Bio CRM [40], an extension of CIDOC CRM for biographical data. WarSampo took first steps towards 3. generation systems as it included some data analytic tools for, e.g., visualizing the casualties of troops and other prosopographical groups, such as officers, on a timeline. The system also presented various statistics pertaining to the fallen soldiers buried in the over 600 Finnish war cemeteries [41].

The idea of integrating data-analysis with semantic faceted search and browsing was developed further in the *Norssi Alumni* portal[23] (online since 2017) using a historical registry of ca 10 000 students of the prominent Finnish high school "Norssi" in 1867–1992. Here faceted search was used for filtering our groups of people and a number of statistical tools and visualizations could be applied to the result set. For example, most common later vocations, work places, or hobbies of the students in different times could be found and studied. The Norssi Alumni system was re-used and developed further in the *U.S. Congress Prosopographer* system[24] (online since 2018) [6] where a registry[25] of all U.S. Congress legislators from the 1st through 115th Congresses (1789–2018) was used. In this case, legislator data could be visualized of maps, and there were specific tabs for comparing statistics and map views of democratic and republican legislators.

The idea of 3. generation biographical systems and using biographies as linked (open) data was fully developed in the system *BiographySampo – Finnish Biographies on the Semantic Web*[26] (online since 2018) [42], a popular web service with tens of thousands of users. The system is based on mining out a large knowledge graph from the ca. 13 100 Finnish national biographies of the Finnish Literature Society, authored by some 940 scholars. The data is interlinked and enriched internally by 16 external data sources and by reasoning, e.g., by inferring family relations [31] and connections of interest between people and places [37]. In addition, a large linguistic knowledge graph of some 120 million triples of the biography texts was created and used for linguistic analyses about the biographies and their authors. For example, it was found that family-related words are widely used in biographies of female Members of Parliament but not for male.

[20] https://kansallisbiografia.fi/english/national-biography
[21] https://www.cidoc-crm.org/
[22] Project: https://seco.cs.aalto.fi/projects/sotasampo/; portal: https://www.sotasampo.fi/
[23] Portal: http://www.norssit.fi/semweb/
[24] Portal: https://semanticcomputing.github.io/congress-legislators/
[25] https://github.com/unitedstates/congress-legislators
[26] Project: https://seco.cs.aalto.fi/projects/biografiasampo/; portal: https://biografiasampo.fi/

A set of data analyses from different perspectives of the BiographySampo dataset is presented in [8], using both the portal user interface and the underlying SPARQL service via the YASGUI editor[27] [43], Google Colab[28], and Jupyter notebooks[29]. The analyses showed, for example, various statistics on charts, graphs and matrices on how the vocations of biographees change in time and correlate between parents and children, what places are mentioned in biographies and when on maps, and visualization of networks of biographees on how they relate to each other based on mentions in the biographies and on family and other relations.

One application perspective of the BiographySampo portal, based on relational search [37], can be seen as an example of a 4. generation system: Here the user first constrains freely people, professions, and places of interest to her/him using faceted search, and the system then finds "interesting" semantic connections between them and creates natural language explanations for the relations found. For example, when the used selects "Italy" from the place facet and "artist" in the profession faceted, one of the answers to the query is "Elin Danielson-Gambogi got the Florence City Award in 1899" based on an event extracted from her biography text and the place ontology telling that Florence is part of Italy.

After BiographySampo, the idea of integrating data publishing with data analysis was reused in the system *WarVictimsSampo 1914–1922* (online since 2019) [9] with data about the 41 500 victims and 1200 battles of the Finnish civil war and kindred wars. In this system, a new tool for interface design, Sampo-UI [3] was utilized. The application included an automatically generated animation on how the deaths in battles spread in Finland as time goes by in 1918.

The latest biographical Sampo system is *AcademySampo* (online since 2021) [31, 11], a biographical in-use LOD service and semantic portal based on 28 000 short biographies of all known Finnish academic people educated in Finland in 1640–1899. The system includes a rich set of data-analytic tools for DH research [10].

4. Sampo Model for Publishing Biographies

Table 3
Sampo Model principles for LOD publishing (P1–P3) and portal logic design (P4–P6)

P1	Support collaborative data creation and publishing
P2	Use a shared open ontology infrastructure
P3	Make clear distinction between the LOD service and the user interface (UI)
P4	Provide multiple perspectives to the same data
P5	Standardize portal usage by a simple filter-analyze two-step cycle
P6	Support data analysis and knowledge discovery in addition to data exploration

The work on biographical applications described above contributed to the development of the so called *Sampo Model* and the *Sampo Series* of semantic portals and LOD services[30] [1]. Based

[27] https://yasgui.triply.cc

[28] https://colab.research.google.com/notebooks/intro.ipynb

[29] https://jupyter.org

[30] See https://seco.cs.aalto.fi/applications/sampo/ for a complete list of "Sampo portals", videos, and further information.

on the six principles listed in Table 3, the model is a kind of consolidated approach for creating LOD services and semantic portals, something that the field of the Semantic Web is arguably still largely missing [44].

Principles P1–P3 can be seen as a foundation for developing LOD services; P4–P6 are related to creating semantic portals.[31] The model is based on the idea of collaborative content creation (P1) from multiple data sources. The data is aggregated from local data silos into a global service, based on a shared ontology [45] and publishing infrastructure (P2). The local data are harmonized and enriched with each other by linking and reasoning. In this model everybody can arguably win, including the data publishers by mutually enriched linked data and by re-using shared publishing infra, and the end users by richer global content and services. The model argues (P3) for the idea of separating the underlying Linked Data service *completely* from the user interface via a SPARQL API. This arguably simplifies the portal architecture and the data service can be opened for data analysis research and application development in Digital Humanities for everybody.

The general idea of principles P4–P6 is to "standardize" the UI logic so that the portals are easier to use for the end users and for the programmers to develop [3]. Principle P4 articulates the idea of providing different thematic *application perspectives* by re-using the data service. The application perspectives can be provided on the landing page of the Sampo portal system or be completely separate applications by third parties. According to P5 the perspectives can be used by a two-step cycle for research: First the focus of interest, the target group, is filtered out using faceted semantic search [46, 47, 48]. Second, the target group is visualized or analyzed by using ready-to-use data analytic tools of the application perspectives. At this point, it is also possible to select a particular member in the target group for a closer look and explore the data by browsing related links. Finally, the Sampo model aims not only at data publishing with search and data exploration [49] but also to data analysis and knowledge discovery with seamlessly integrated tooling for finding, analyzing, and even solving research problems in interactive ways (P6). The Sampo model principles are compatible with the FAIR principles for creating Findable, Accessible, Interoperable, and Re-usable data[32], but were developed in the context of publishing and using LOD.

The Sampo model has been used in the biographical Sampo systems listed in Table 1. They make use of several data sources that enrich each other (P1). A set of shared ontologies have been used for harmonizing and amalgamating the datasets (P2). The aggregated data is then published as a separate LOD services including a SPARQL endpoint (3). On top of the data service different applications were created using only SPARQL for accessing the data. For example, the WarSampo portal initially contained six application perspectives and three more were added later. A Sampo system can also be re-used by external applications, such as other Sampos, leading to a kind of "Sampo cloud". For example, the *WarMemoirSampo* system[33] [50] for publishing video memoirs of the Second World War veterans is based on and enriched by the WarSampo data infrastructure and portal pages.

[31] The model is called "Sampo" according to the Finnish epic Kalevala, where Sampo is a mythical machine giving riches and fortune to its holder, a kind of ancient metaphor of technology according to the most common interpretation of the concept.

[32] https://www.go-fair.org/fair-principles/

[33] Portal: https://sotamuistot.arkisto.fi; Project home: https://seco.cs.aalto.fi/projects/war-memoirs/

For publishing the LOD services, the Linked Data Finland service LDF.fi[34] [51] has been used. In LDF.fi, the 5-star model[35] of Tim Berners-Lee is extended to a 7-star model. The 6th star is given to a data publication if it includes not only the 5-star data but also the schemas of the data with documentation. This makes re-use of data easier. The 7th star is given to a data publication, if the publication includes some kind of evaluation that the data actually conforms to the provided schemas using, e.g., SHACL[36] or ShEx[37] [52]. The idea here is to encourage publishers to publish high quality LOD, which is a severe issue on the Semantic Web.

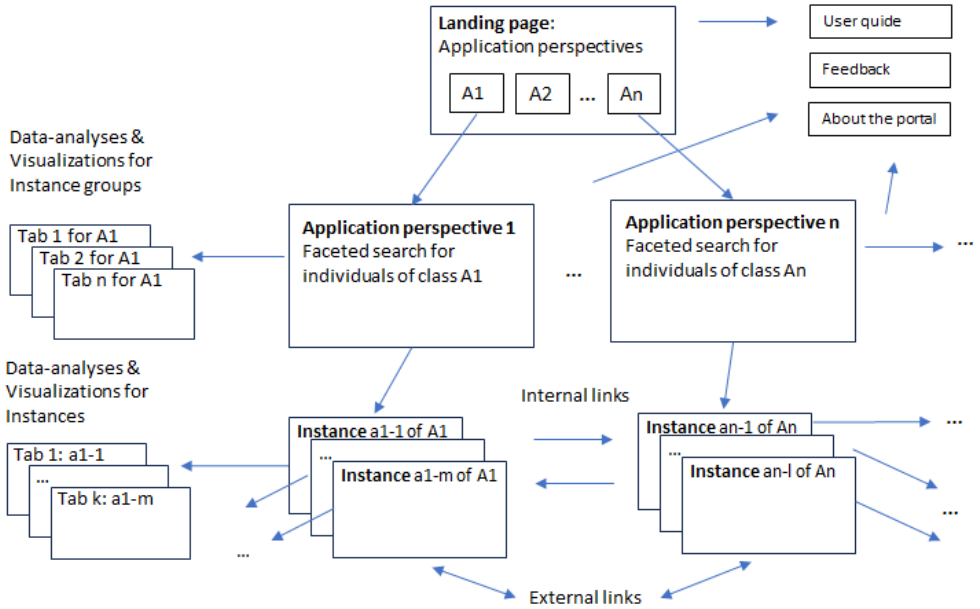

Figure 1: Navigational page structure of a Sampo portal based on Sampo-UI

The Sampo model principles P4–P6 are used for designing the portal user interfaces: the idea is to "standardize" the UI logic of Sampo portals to be created on top of a LD service SPARQL endpoint. The goal is to make the portals easier to use and implement. Fig. 1 illustrates the navigational structure of a Sampo portal. The user first lands on the *landing page* with several *application perspectives* to the data. The landing page that introduces the application perspectives as clickable boxes; by selecting one the corresponding perspective is opened. Each perspective typically provides a faceted search engine for filtering out a target group of individual instances of a class, such as people, places, or events. The search result can be visualized and studied on separate tabs. The default is to list results as a table, but the results can also be studied by data analytic tools on maps, using statistical charts, on timelines, or as networks for network analysis. Each individual in the system, say a person or a place, has a "home page" on which data related to

[34] https://ldf.fi
[35] https://www.w3.org/community/webize/2014/01/17/what-is-5-star-linked-data/
[36] https://www.w3.org/TR/shacl/
[37] https://shex.io/

it is automatically aggregated for providing a rich contextualized representation of the individual. Also on the home page, a set of tabs can be provided as data-analytic views of the individual. For example, for a person home page an egocentric network of related other persons can be shown or events in which the person participated in different roles can be visualized on a map or timeline.

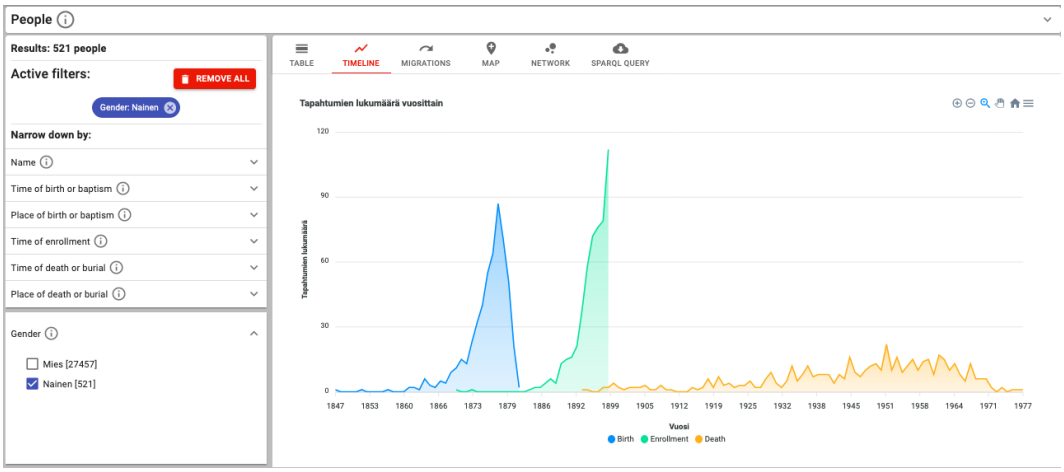

Figure 2: The annual births, enrollments, and deaths of female students on the TIMELINE tab. The target group has been filtered out by the facets shown on the left. Six alternative tabs for showing and analyzing the group can be selected on the top.

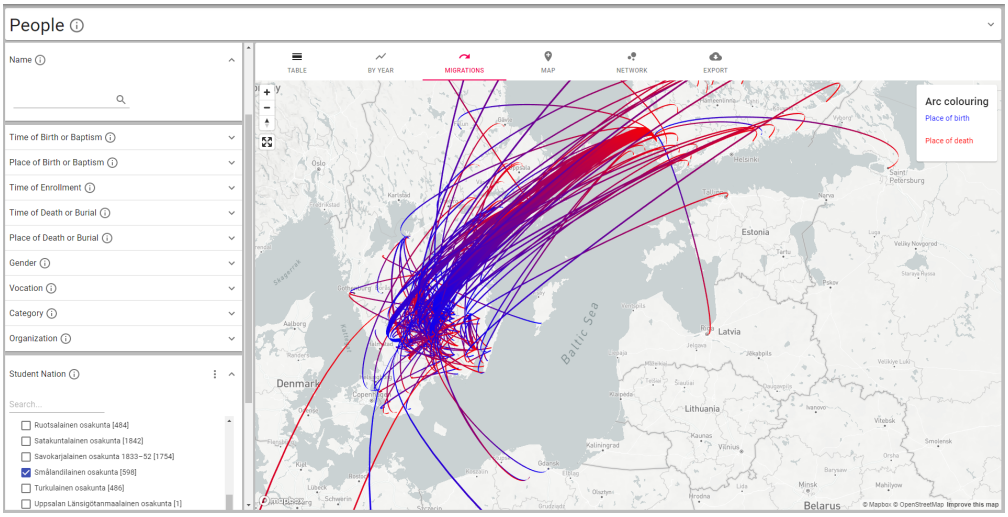

Figure 3: Using AcademySampo for prosopographical research by visualizing life charts of the members of the Student Nation of Småland as arcs from the place of birth (blue end of the arc) to the place of death (red end)

In practice, the UIs of the biographical Sampos have been implemented using the tools SPARQL

Figure 4: The MAP tab visualizes the lifetime places mentioned in about 175 000 events.

Faceter [2] (in WarSampo, Norssi Alumni, U.S. Congress Prosopographer, and BiographySampo) and since 2018 with Sampo-UI [3] (in WarWictimSampo and AcademySampo). For example, in the People application perspective of AcademySampo [10], prosopographical analysis tools are available in the TIMELINE, MIGRATION, MAP, and NETWORK tabs in addition to the default TABLE tab listing the search results. The TIMELINE tab shows the annual births, university enrollments, and deaths of the filtered people. For instance, Fig. 2 shows the charts for all 521 female university students in Finland in 1640–1899. The MIGRATION tab (Fig. 3) visualizes the mobility and immigration of students (597) of the Swedish Småland Student Nation with arcs depicting the life cycles. The blue end of the arc indicates the place of birth and the red the place of death, which is most often in the territory of present-day Finland, and the thickness of the arc reflects the number of people associated with the arc. If a person was born and died in the same place, the arc is not displayed. By clicking on the arc, one will find related links to people's home pages. The MAP tab (Fig. 4) shows the approx. 3000 locations to which students are connected by approximately 175 000 events. For example, clicking on a marker in Ireland finds two related people, the other being the famous Johan Gadolin (1760–1855), who later discovered a new element, Yttrium. Finally, the NETWORK tab allows to explore the internal academic network of a group of people specified by the facet selections, for example, the teacher-student network of 1480 male students born in Helsinki.

5. Discussion

This paper presented a vision on how publishing and using biographies has evolved from engravings and written texts (1. generation) to Web-based publishing (2. generation), publishing biographies as Linked Open Data with seamlessly integrated data-analytic tools (3. generation), and finally to knowledge discovery-based systems (4. generation) based on AI. As an attempt to realize this vision, a series of biographical Sampo systems in use in Finland were created. These systems make use of the Sampo model, a set of principles for creating LOD services

and biographical portals on top of them. Our empirical experiences suggests that the model is feasible both from the end-user's and data publisher's points of view. More information about the applications as well and the underlying technical challenges and solutions can be found in the papers and web addresses given in relation to the applications.

During the work also several challenges of using linked data and the Sampo model have been encountered. Using explicit ontologies and linked data sets more demands on data quality than before. Any problems of data modelling or quality are highlighted in the user interfaces and data analyses. In most cases automatic annotation and linking had to be used for knowledge extraction in biographies, which lowers data quality, but on the other hand, manual annotations are costly and do not scale up. The Sampo model also requires more collaboration between the data publishers regarding interoperability, which complicates work. Integration of semantic portals with legacy systems can be a practical challenge in many organizations as well as sustainable maintenance of interlinked knowledge graphs [53]. From the end user, more source criticism[38] and understanding the characteristics and limitations of data are needed [54, 55]. However, the challenges in my mind seem to be smaller than the benefits and potential of utilizing linked open data in publishing and using biographical contents on the Web.

Acknowledgements

Our work on biographies is partly supported by the EU project InTaVia: In/Tangible European Heritage[39], and is related to the EU COST action Nexus Linguarum[40] on linguistic data science. Thanks to the Finnish Cultural Foundation for an Eminentia grant and to CSC – IT Center for Science for providing computational resources.

References

[1] E. Hyvönen, Digital humanities on the semantic web: Sampo model and portal series, Semantic Web – Interoperability, Usability, Applicability 14 (2023) 729–744. doi:10.3233/SW-223034.

[2] M. Koho, E. Heino, E. Hyvönen, 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. URL: http://ceur-ws.org/Vol-1615/semdevPaper5.pdf.

[3] E. Ikkala, E. Hyvönen, H. Rantala, M. Koho, Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces, Semantic Web – Interoperability, Usability, Applicability 13 (2022) 69–84. doi:10.3233/SW-210428.

[4] E. Hyvönen, E. Heino, P. Leskinen, E. Ikkala, M. Koho, M. Tamper, J. Tuominen, E. Mäkelä, WarSampo data service and semantic portal for publishing linked open data about the Second World War history, in: H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S. P. Ponzetto,

[38] https://ranke2.uni.lu/define-dsc/#%20,%20Universit%C3%A9%20du%20Luxembourg

[39] https://intavia.eu/

[40] https://nexuslinguarum.eu/the-action

C. Lange (Eds.), The Semantic Web – Latest Advances and New Domains (ESWC 2016), Springer, 2016, pp. 758–773. doi:10.1007/978-3-319-34129-3_46.

[5] E. Hyvönen, P. Leskinen, E. Heino, J. Tuominen, L. Sirola, Reassembling and enriching the life stories in printed biographical registers: Norssi high school alumni on the Semantic Web, in: Proceedings, Language, Technology and Knowledge (LDK 2017), Springer, 2017, pp. 113–119. doi:10.1007/978-3-319-59888-8_9.

[6] G. Miyakita, P. Leskinen, E. Hyvönen, Using linked data for prosopographical research of historical persons: Case U.S. Congress Legislators, in: Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection: 7th International Conference, EuroMed 2018, Nicosia, Cyprus, October 29-November 3, 2018, Proceedings. Part II, volume 11197 LNCS, Springer International Publishing, 2018, pp. 150–162. doi:10.1007/978-3-030-01765-1_18.

[7] E. Hyvönen, P. Leskinen, M. Tamper, H. Rantala, E. Ikkala, J. Tuominen, K. Keravuori, BiographySampo – publishing and enriching biographies on the semantic web for digital humanities research, in: The Semantic Web - 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2-6, 2019, Proceedings, volume 11503 LNCS, Springer International Publishing, 2019, pp. 574–589. doi:10.1007/978-3-030-21348-0{_}37.

[8] M. Tamper, P. Leskinen, E. Hyvönen, R. Valjus, K. Keravuori, Analyzing biography collection historiographically as linked data: Case National Biography of Finland, Semantic Web – Interoperability, Usability, Applicability 14 (2023) 385–419. URL: https://doi.org/10.3233/SW-222887.

[9] H. Rantala, I. Jokipii, E. Ikkala, E. Hyvönen, WarVictimSampo 1914–1922: a national war memorial on the semantic web for digital humanities research and applications, ACM Journal on Computing and Cultural Heritage 15 (2022). URL: https://doi.org/10.1145/3477606. doi:0.1145/3477606.

[10] P. Leskinen, H. Rantala, E. Hyvönen, Analyzing the lives of finnish academic people 1640–1899 in Nordic and Baltic countries: AcademySampo data service and portal, in: 6th Digital Humanities in Nordic and Baltic Countries Conference, Proceedings, CEUR Workshop Proceedings, 2022. URL: https://seco.cs.aalto.fi/publications/2022/leskinen-et-al-academysampo-dhnb-2022.pdf, forth-coming.

[11] P. Leskinen, E. Hyvönen, Reconciling and using historical person registers as linked open data in the AcademySampo knowledge graph, in: The Semantic Web – ISWC 2021. 20th International Semantic Web Conference, ISWC 2021,Proceedings, Springer, 2021, pp. 714–730. doi:10.1007/978-3-030-88361-4_42.

[12] T. Keith, Changing conceptions of National Biography, Cambridge University Press, 2005. doi:10.1017/cbo9780511497582.

[13] M. Klinge (Ed.), Suomen kansallisbiografia 1–10, Suomalaisen Kirjallisuuden Seura, Helsinki, Finland, 2003–2007.

[14] C. N. Warren, Historiography's two voices: Data infrastructure and history at scale in the oxford dictionary of national biography (ODNB), Journal of Cultural Analytics 1 (2018) 1–31. doi:10.22148/16.028.

[15] Ú. Bhreathnach, C. Burke, J. M. Fhinn, G. Ó. Cleircín, B. Ó. Raghallaigh, A quantitative analysis of biographical data from Ainm, the Irish-language biographical database, in:

BD2019 Biographical Data in a Digital World 2019. Proceedings of the Third Conference on Biographical Data in a Digital World 2019, volume 3152, CEUR Workshop Proceedings, 2019. URL: http://ceur-ws.org/Vol-3152/.

[16] A. Fokkens, S. ter Braake, N. Ockeloen, P. Vossen, S. Legêne, G. Schreiber, V. de Boer, BiographyNet: Extracting Relations Between People and Events, New Academic Press, Berlin, Germany, 2017, pp. 193–224.

[17] M. Schlögl, K. Lejtovicz, A prosopographical information system (APIS), in: Proceedings of the Second Conference on Biographical Data in a Digital World 2017 Linz, Austria, November 6-7, 2017, volume 2119, CEUR Workshop Proceedings, 2018. URL: http://ceur-ws.org/Vol-2119/.

[18] A. Jatowt, D. Kawai, K. Tanaka, Time-focused analysis of connectivity and popularity of historical persons in Wikipedia, International Journal on Digital Libraries 20 (2019) 287–305. doi:10.1007/s00799-018-0231-4.

[19] D. Metilli, V. Bartalesi, C. Meghini, A Wikidata-based tool for building and visualising narratives, International Journal on Digital Libraries 20 (2019) 417–432. doi:10.1007/s00799-019-00266-3.

[20] S. ter Braake, A. Fokkens, R. Sluijter, T. Declerck, E. Wandl-Vogt (Eds.), BD2015 Biographical Data in a Digital World 2015, volume 1399, CEUR Workshop Proceedings, 2015. URL: http://ceur-ws.org/Vol-1399/.

[21] A. Fokkens, S. ter Braake, R. Sluijter, P. Arthur, E. Wandl-Vogt (Eds.), BD-2017 Biographical Data in a Digital World 2017, volume 2119, CEUR Workshop Proceedings, 2017. URL: http://ceur-ws.org/Vol-2119/.

[22] R. Larson, Bringing lives to light: Biography in context. Final project report, 2010. URL: http://metadata.berkeley.edu/Biography_Final_Report.pdf, University of Berkeley.

[23] C. Warren, D. Shore, J. Otis, L. Wang, M. Finegold, C. Shalizi, Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks, Digital Humanities Quarterly 10 (2016) 1–16.

[24] A. Langmead, J. Otis, C. Warren, S. Weingart, L. Zilinski, Towards Interoperable Network Ontologies for the Digital Humanities, International Journal of Humanities and Arts Computing 10 (2016). doi:http://dx.doi.org/10.3366/ijhac.2016.0157.

[25] M. Tamper, E. Hyvönen, P. Leskinen, Visualizing and analyzing networks of named entities in biographical dictionaries for Digital Humanities research, in: Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICling 2019), Springer, 2021. Preprint: https://seco.cs.aalto.fi/publications/2021/tamper-et-al-cicling-2021.pdf.

[26] T. Heath, C. Bizer, Linked Data: Evolving the Web into a Global Data Space (1st edition), Morgan & Claypool, Palo Alto, California, 2011. doi:10.2200/S00334ED1V01Y201102WBE001.

[27] J. L. Martinez-Rodriguez, A. Hogan, I. Lopez-Arevalo, Information extraction meets the semantic web: A survey, Semantic Web – Interoperability, Usability, Applicability 11 (2020) 255–335. doi:10.3233/SW-180333.

[28] A. Gangemi, V. Presutti, D. R. Recupero, A. G. Nuzzolese, F. Draicchio, M. Mongiovì, Semantic web machine reading with FRED, Semantic Web – Interoperability, Usability, Applicability 8 (2017) 873–893. doi:10.3233/sw-160240.

[29] M. C. Pattuelli, M. Miller, L. Lange, H. K. Thorsen, Linked Jazz 52nd Street: A LOD Crowdsourcing Tool to Reveal Connections among Jazz Artists., in: 8th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2013, Lincoln, NE, USA, July 16-19, 2013, Conference Abstracts, Alliance of Digital Humanities Organizations (ADHO), 2013, pp. 337–339.

[30] M. Rospocher, M. van Erp, P. Vossen, A. Fokkens, I. Aldabe, G. Rigau, A. Soroa, T. Ploeger, T. Bogaard, Building event-centric knowledge graphs from news, Web Semantics: Science, Services and Agents on the WWW 37 (2016) 132–151. doi:10.2139/ssrn.3199233.

[31] P. Leskinen, E. Hyvönen, Linked open data service about historical Finnish academic people in 1640–1899, in: DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference, volume 2612, CEUR Workshop Proceedings, 2020, pp. 284–292. URL: http://ceur-ws.org/Vol-2612/short14.pdf.

[32] M. Tamper, P. Leskinen, K. Apajalahti, E. Hyvönen, 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, 2018, pp. 125–137. doi:10.1007/978-3-030-01762-0_11.

[33] Á. Z. Bernád, M. Kaiser, The biographical formula: Types and dimensions of biographical networks, in: Proceedings of the Second Conference on Biographical Data in a Digital World 2017 Linz, Austria, November 6-7, 2017., volume 2119, CEUR Workshop Proceedings, 2018. URL: http://ceur-ws.org/Vol-2119/.

[34] V. Gunter, S. Matthias, G. Vogeler, Data exchange in practice: Towards a prosopographical api (preprint), in: Proceedings of the Third Conference on Biographical Data in a Digital World (BD 2019), Varna, Bulgaria, 2019.

[35] E. Hyvönen, Using the semantic web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery, Semantic Web – Interoperability, Usability, Applicability 11 (2020) 187–193. doi:10.3233/SW-190386.

[36] E. Gardiner, R. G. Musto, The Digital Humanities: A Primer for Students and Scholars, Cambridge University Press, New York, NY, USA, 2015. doi:10.1017/CBO9781139003865.

[37] E. Hyvönen, H. Rantala, Knowledge-based relational search in cultural heritage linked data, Digital Scholarship in the Humanities (DSH), Oxford University Press 36 (2021) 55–64. doi:10.1093/llc/fqab042.

[38] E. Hyvönen, M. Alonen, E. Ikkala, E. Mäkelä, Life stories as event-based linked data: Case semantic National Biography, in: Proceedings of the ISWC 2014 Posters & Demonstrations Track, a track within the 13th International Semantic Web Conference (ISWC 2014) Riva del Garda, Italy, October 21, 2014., volume 1272, CEUR Workshop Proceedings, 2014, pp. 1–4. URL: http://ceur-ws.org/Vol-1272/paper_5.pdf.

[39] M. Doerr, The CIDOC CRM – an ontological approach to semantic interoperability of metadata, AI Magazine 24 (2003) 75–92.

[40] J. Tuominen, E. Hyvönen, P. Leskinen, 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 Linz, Austria, November 6-7, 2017, volume 2119, CEUR Workshop Proceedings, 2018. URL: http://ceur-ws.org/Vol-2119/.

[41] E. Ikkala, M. Koho, E. Heino, P. Leskinen, E. Hyvönen, T. Ahoranta, Prosopographical views to finnish ww2 casualties through cemeteries and linked open data, in: Proceedings of the Workshop on Humanities in the Semantic Web (WHiSe II), volume 2014, CEUR Workshop Proceedings, 2017. URL: http://ceur-ws.org/Vol-2014/.

[42] E. Hyvönen, P. Leskinen, M. Tamper, H. Rantala, E. Ikkala, J. Tuominen, K. Keravuori, BiographySampo – Publishing and enriching biographies on the Semantic Web for digital humanities research, in: The Semantic Web. 16th International Conference, ESWC 2019, Springer, 2019, pp. 574–589. doi:10.1007/978-3-030-21348-0_37.

[43] L. Rietveld, R. Hoekstra, The YASGUI family of SPARQL clients, Semantic Web – Interoperability, Usability, Applicability 8 (2017) 373–383. doi:10.3233/SW-150197.

[44] P. Hitzler, A review of the semantic web field, Commun. ACM 64 (2021) 76–83. doi:10.1145/3397512.

[45] S. Staab, R. Studer (Eds.), Handbook on Ontologies (2nd Edition), Springer, 2009.

[46] E. Hyvönen, S. Saarela, K. Viljanen, Application of ontology-based techniques to view-based semantic search and browsing, in: Proceedings of the First European Semantic Web Symposium, Springer, 2004. doi:10.1007/978-3-540-25956-5_7.

[47] D. Tunkelang, Faceted search, Morgan & Claypool, Palo Alto, California, 2009. doi:10.2200/S00190ED1V01Y200904ICR005.

[48] Y. Tzitzikas, N. Manolis, P. Papadakos, Faceted exploration of RDF/S datasets: a survey, Journal of Intelligent Information Systems 48 (2017) 329–364. doi:10.1007/s10844-016-0413-8.

[49] G. Marchionini, Exploratory search: from finding to understanding, Communications of the ACM 49 (2006) 41–46. doi:10.1145/1121949.1121979.

[50] E. Hyvönen, E. Ikkala, M. Koho, R. Leal, H. Rantala, M. Tamper, How to search and contextualize scenes inside videos for enriched watching experience: Case stories of the second world war veterans, in: Proceedings of the 19th Extended Semantic Web Conference (ESWC 2022), Poster and Demo Papers, 2022. URL: https://seco.cs.aalto.fi/publications/2022/hyvonen-et-al-wms-2022.pdf, forth-coming.

[51] E. Hyvönen, J. Tuominen, M. Alonen, E. Mäkelä, Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets, in: The Semantic Web: ESWC 2014 Satellite Events, Springer, 2014, pp. 226–230. doi:10.1007/978-3-319-11955-7_24.

[52] J. E. Labra Gayo, E. Prud'hommeaux, I. Boneva, D. Kontokostas, Validating RDF Data, volume 7 of *Synthesis Lectures on the Semantic Web: Theory and Technology*, Morgan & Claypool Publishers LLC, 2017. URL: https://doi.org/10.2200/s00786ed1v01y201707wbe016. doi:10.2200/s00786ed1v01y201707wbe016.

[53] C. Gutierrez, J. F. Sequeda, Knowledge graphs, Communications of the ACM 64 (2021) 96–104. doi:10.1145/3418294.

[54] T. Koltay, Data literacy for researchers and data librarians, Journal of Librarianship and Information Science 49 (2015) 3–14. doi:10.1177/0961000615616450.

[55] E. Mäkelä, K. Lagus, L. Lahti, T. Säily, M. Tolonen, M. Hämäläinen, S. Kaislaniemi, T. Nevalainen, Wrangling with non-standard data, in: Proceedings of the Digital Humanities in the Nordic Countries 5th Conference, CEUR Workshop Proceedings, CEUR-WS.org, Germany, 2020, pp. 81–96.