

Analyzing the Lives of Finnish Academic People 1640–1899 in Nordic and Baltic Countries: AcademySampo Data Service and Portal

Petri Leskinen^{1,2}, Eero Hyvönen^{1,2} and Heikki Rantala^{1,2}

¹*Semantic Computing Research Group (SeCo), Aalto University, Finland*

²*Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland*

Abstract

This paper shows how the newly published Linked Open Data (LOD) service and semantic portal “ACADEMYSAMPO – Finnish Academic People 1640–1899” can be used for Digital Humanities (DH) research. The original primary data, based on some ten man-years of digitization work, covers a significant part of the Finnish university history based on the student registries in 1640–1852 and 1853–1899. They contain biographical descriptions of 28 000 students of the University of Helsinki, originally the Royal Academy of Turku. ACADEMYSAMPO also sheds light to the academic history of Sweden and Baltic countries through their shared history with Finland in the larger Swedish Empire. The Finnish student registries have been widely used by genealogists and historians by close reading. We argue that unprecedented new possibilities for DH research are now enabled by using ACADEMYSAMPO: the underlying knowledge graph can be accessed and analyzed using Semantic Web technologies and tools and with the ready-to-use data-analytic tools of the portal. Examples of data-analysis are presented by using the ACADEMYSAMPO system for studying migrations of students in Finland, Sweden, Russia, and Estonia, history of student nations, inheritance of vocations and social classes, lengths of family lines of students, and network analyses of students. Related analyses have been made before using biographical dictionaries but not for academic history and student registries.

Keywords

Linked Data, Data Analysis, Digital Humanities, Network Analysis, Cultural Heritage

1. Introduction

ACADEMYSAMPO¹ [1, 2] consists of two parts: 1) a portal² and 2) a LOD service³ published on the Linked Data Finland platform [3]. The ACADEMYSAMPO Portal provides intelligent capabilities for searching and browsing with seamlessly integrated data analytical tools and visualizations for biographical and prosopographical [4] research using statistics, networks, timelines, and maps. Using of the portal does not require special IT skills. The open Application Programming Interfaces (API) of the LOD service and its SPARQL endpoint, in turn, provide an easy-to-use

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022.

✉ petri.leskinen@aalto.fi (P. Leskinen)

📞 0000-0003-2327-6942 (P. Leskinen); 0000-0003-1695-5840 (E. Hyvönen); 0000-0002-4716-6564 (H. Rantala)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹Project homepage: <https://seco.cs.aalto.fi/projects/yo-matrikkelit/>

²Portal was opened February 2, 2021 at <https://akatemiasampo.fi/en/>

³The LOD service is available at <https://ldf.fi/dataset/yoma>

opportunity to implement new data analyses for DH researchers with some experience in the SPARQL query language⁴ and programming. For example, the YASGUI editor [5], Jupyter⁵, Google Colab notebooks⁶, and Python scripts can be used.

Based on the Sampo model [6], the portal was implemented using the Sampo-UI framework [7] as an example of using the data service in application development. ACADEMYSAMPO is part of the Sampo portal series⁷ and uses the Linked Open Data Infrastructure for Digital Humanities in Finland (LODI4DH)⁸ [8], a part of the Finnish FIN-CLARIAH infrastructure initiative⁹.

This paper describes how the ACADEMYSAMPO portal and LOD service can be used for DH research. In Section 2 the LOD underlying the system is first described. After this using the portal (Section 3) and the underlying SPARQL endpoint (Section 4) is discussed and demonstrated. In conclusion (Section 5), related works are discussed as well as the need for data literacy in using systems like ACADEMYSAMPO.

2. Primary Data and Knowledge Graph

ACADEMYSAMPO's data form an extensive knowledge graph that has been produced algorithmically from the digitized student registers of the Royal Academy of Turku and the University of Helsinki in 1640–1852 and 1853–1899¹⁰ by extracting information from the texts and database structures. The data has been enriched by linking it both internally by artificial intelligence-based reasoning, and externally to other open datasets. The student registers describe all people who have received academic education in Finland in 1640–1899, as there were no other universities in Finland at that time. The descriptions of students tell not only about their studies, but also about their career after studies and relatives, as well as references to the literature. The original register of the Royal Academy of Turku was destroyed in the Great Fire of Turku in 1827, but it was reconstructed in the late 19th century by Vilhelm Lagus. The register was supplemented in the 20th century from various sources, and in the end the information was edited by Yrjö Kotivuori and Veli-Matti Autio in an effort of ca. ten man years.

The registers 1640–1852 and 1853–1899 are digitized and provided by different authors, and their tabular CSV data differ to some extent. The source information found in the table of records 1640–1852 includes, in addition to some technical information in the database: 1) the person's registration number, 2) HTML text showing the person's name, places and times of birth and death, parents, career events, relatives, students, references and 3) the date the record was created. If the person mentioned in the register 1640–1852 is found in either of the registers, a HTML link is manually created connecting this mention to the person's page using the registration number. However, in the register 1853–1899 there are no such links, and the references have been interpreted computationally. In addition, supplementary textual information about a person may be available in other registers. For example, Johan Ludvig

⁴<https://www.w3.org/TR/sparql11-query/>

⁵Jupyter Project and Tool: <https://jupyter.org>

⁶Google Colab: <https://colab.research.google.com/notebooks/intro.ipynb#recent=true>

⁷See: <https://seco.cs.aalto.fi/applications/sampo/>

⁸LODI4DH initiative: <https://seco.cs.aalto.fi/projects/lodi4dh/>

⁹<https://seco.cs.aalto.fi/projects/fin-clariah/>

¹⁰Student Registers, University of Helsinki: <https://www.helsinki.fi/fi/yliopisto/ylioppilasmatrikkelit-1640-1907>

Runeberg has further information in the registers of Lagus and Carpelan.

The primary data used by in creating ACADEMYSAMPO was therefore mainly text in HTML format without structured metadata, such as places or times of birth, vocation, etc. A major technical challenge in creating the linked data was to unambiguously identify the entities and events mentioned in the text, such as marriages, rewards and promotions, and key concepts, such as vocations. A specific challenge in extracting information was to distinguish between people with the same name, to reason their gender by name, and to infer various relationships, such as little cousin, through other relationships.

The data of the ACADEMYSAMPO was converted into Linked Data [9]¹¹ by structuring the text descriptions of the Student register 1640–1852 for about 9500 people and the register 1853–1899 for about 18 450 people. This was done by identifying, through regular expressions, basic biographical information about students, their 47 000 relatives, 120 000 interpersonal relationships, 3000 historical places, 10 000 vocations, and 4000 academic student–teacher relationships. The “semantic glue” of the knowledge graph are the events related to the professional and family life of the 175 000 people identified in the texts, which link the people and organizations involved in different roles with places and times according to the CIDOC CRM¹² ontology and ISO standard. The data has been enriched by linkage to external databases, such as the Finnish National Biography and other biographies of the Finnish Literature Society available as LOD in the BiographySampo system [10] and Wikidata¹³, and by inferring relationships between people [11, 12]. The public data service is available at the Linked Data Finland for accessing and utilizing the data in research and application development, such as the ACADEMYSAMPO portal.

3. Using AcademySampo Portal for Data Analysis

In accordance with the principles of the Sampo model and Sampo-UI programming framework, the portal offers four application perspectives to the Student Register materials shown in Figure 1: People, Places, Vocations, and Student Nations. Clicking on the corresponding icon opens views for searching, exploring, and studying people, places, vocations, and student nations, either as individuals (biography) or as groups of individuals (prosopography) through the methods of DH. In the following, the main functionalities of the portal are described.

(1) People Perspective In the “People” perspective one can search for a specific individual or group of people for prosopographical analysis using ontology-based faceted search depicted in Figure 2. The 13 search facets are on the left and the results fill the rest of the page on the right. By default, the people in the result set are ordered by their degree of networking, e.g., by the number of links to external databases, with famous students such as poet Elias Lönnrot, president of Finland Juho Kusti Paasikivi, and poet Johan Ludvig Runeberg among the top positions. The results can be visualized and analyzed by selecting any of the six tabs available, and sorted column-wise by, e.g., date of birth or place of death.

In faceted search, the result set is filtered by making selections from (possibly) hierarchical facets. As an example, two facets have been opened in Figure 3. The upper one lists the

¹¹W3C Linked Data: <https://www.w3.org/standards/semanticweb/data>

¹²CIDOC CRM: <http://cidoc-crm.org>

¹³Wikidata: <https://wikidata.org>

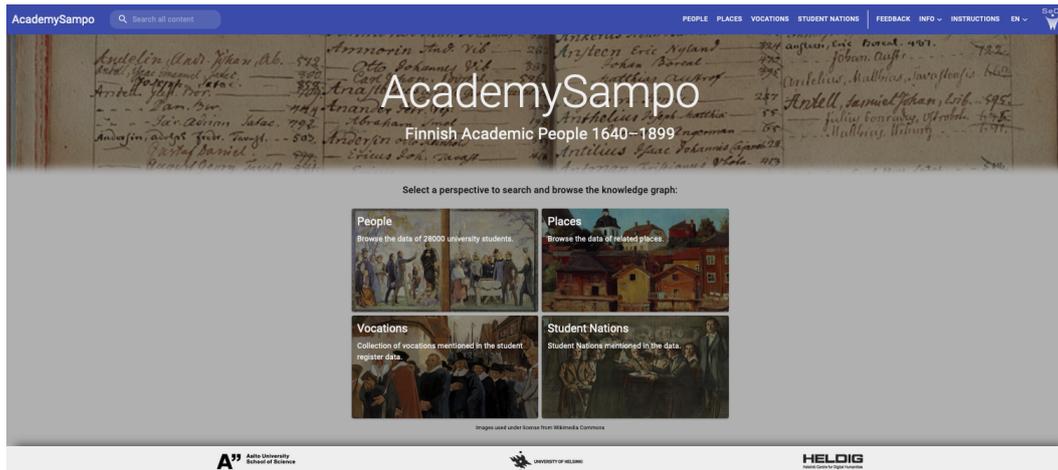


Figure 1: The landing page of the ACADEMYSAMPO portal with four application perspectives (views)

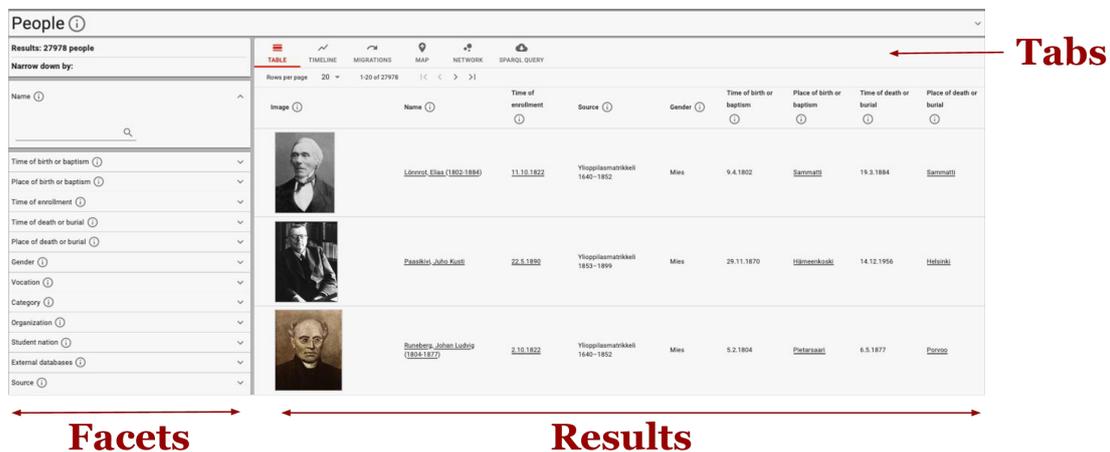


Figure 2: Faceted search using 13 facets for people with six tabs for data analyses of results

organizations associated with the people from which the “Regiment Horn” has been selected, and the lower the student nations in which the people associated with the Regiment Horn can be found. Of these, the “Ostrobothnian Student Nation” has been selected, in which case all four people associated with the Horn Regiment and the Ostrobothnian Student Nation have been found through two facet selections. It should be noted that in reality there may be other such students but the search will not find them if, for some reason, it is not mentioned in ACADEMYSAMPO material (historical data is often incomplete) or if, for some reason, the used algorithms have not been able to separate all the information from the text. All searching and data analysis are, of course, limited to the available data.

Each person in the dataset has a separate instance page or “homepage” showing his or her

People

Organization: Horn

Faceted search results:

- Hel싱in tullikamari [4]
- Hel싱in yleinen sairaala [3]
- Hel싱in yliopisto [30]
- Hel싱in yliopiston kirjasto [1]
- Hornin rykmentti [4]**
- HuVuosokoulut [4]
- Hämeen läänin [14]
- Hämeen läänin jalkaväkirykmentti (jääkripataljoona) [1]
- Hämeen läänin jalkaväkirykmentti [4]
- Hämeen lääninohjaaja [1]
- Itämaannon lääniesikunta [1]

Student nation:

- Uusikoulu [4]
- Borealisen osakunta [1]
- Pohjalaisen osakunta [4]**
- Småländiläinen osakunta [1]
- Vipurilainen osakunta [1]

Name	Time of enrollment	Source	Gender	Time of birth or baptism	Place of birth or baptism	Time of death or burial	Place of death or burial	Vocation	Student nation
Ferdinand Gabriel (1837-1921)	10.10.1855	Ylioppilasmerkki 1640-1852	Mies	1837		8.10.1921	Porvoo	3. alumnus Johan Sironen 4. Hoivakassa (Helsingin yliopisto) 5. Helsingin yliopiston kirjasto	Pohjalainen osakunta
Carlénius Gabriel (1827)	10.10.1843	Ylioppilasmerkki 1640-1852	Mies			1872	II	1. johtokunta 2. Helsingin yliopiston kirjasto 3. Helsingin yliopiston kirjasto 4. Helsingin yliopiston kirjasto 5. Helsingin yliopiston kirjasto	Pohjalainen osakunta
Mattias Jacob (1849)	10.10.1842	Ylioppilasmerkki 1640-1852	Mies			1868	Malmi	1. Helsingin yliopiston kirjasto 2. Helsingin yliopiston kirjasto	Pohjalainen osakunta
John Jacob (1827)	10.10.1842	Ylioppilasmerkki 1640-1852	Mies		Malmi	1872		1. Helsingin yliopiston kirjasto 2. Helsingin yliopiston kirjasto 3. Helsingin yliopiston kirjasto 4. Helsingin yliopiston kirjasto	Pohjalainen osakunta

Figure 3: Searching for people using facet selections; people associated with the Regiment Horn and the Ostrobothnian Student Nation are sought here

Person: Runeberg, Johan Ludvig (1804-1877)

Navigation tabs: TABLE, FAMILY RELATIONS, ACADEMIC RELATIONS, CONNECTIONS, RELATIONS, SPREADSHEET

Image:

Name: Runeberg, Johan Ludvig (1804-1877)

Time of enrollment: 2.10.1822

Alternative names: -

Entry text: 2.10.1822 Johan Ludvig Runeberg (1804-1877). * Pietarsaareissa 5.2.1804. Vht. pietarsaarelainen merikapteeni Loventz Unik Runeberg (1806) (p. 1791, t. 1828) ja Anna Maria Malm. Oulun triviaalkoulun oppilas 6.3.1813 (cl. i). 1814 (avg.). Vaasan triviaalkoulun oppilas 30.1.1815 - 24.7.1822. Pääsykoulutieto 30.9.1822. Ylioppilas Turussa 2.10.1822. Pohjalaisen osakunnan jäsen 5.10.1822. 1822 Johan Ludvig Runeberg, die 5 October. Natus die 4 Februarii anno MDCCCXXII (Pronomen Phisus Doctor anno 1827. Eloquentie Docens 1829. Lector Gymnasii Borgönensis 1837. / De stella polari eques 1844. Professor honorarius nominatus 1845. Utgar' Doktor' 1829. 'Eigalyttarne' s.d. Doktor Zaba h. 1832. 'Helsing' 1836. 'Nobesche' 1841. 'Doktor' 3-dgr. h. 1843. 'Kung' 1844. 'Sankt Olofs Sijoner' 1848. Respondenti 17.6.1825 pro exercitio, pr. Anders Johan Lager (1806). Sigendiattitesei 6.10.1826, pr. Anders Johan Lager (1806). FK 13.6.1827. Respondenti 23.6.1827 pro gradu, pr. Carl Reinhold Sahlberg (1825). FM 10.7.1827. Presea 16.6.1830 pro venia docendi. Presea 16.6.1830 pro venia docendi. Presea 30.11.1833 pro materia. Visitty papiksi Porvoon hiopikoulussa 19.12.1838. TT h.c. 28.5.1857 (käs. mäs. m. nimittämänä 14.5.1857). - Alkuaikainen yliopiston konsistorin amanuenssi 1830-34, virkavapaa 1831 ja 1832-33, kauppakoulun dosentti 1830, ero 1836. Samalla Helsingin kirkkohölysen opettaja 1831-36. Samanaikaisesti vuodesta 1832. Porvoon lukion Rooman kirjallisuuden lehtori 1837, Kreikan kirjallisuuden lehtori 1842, ero 1857. Samalla Porvoon yksityisen esikoulun opettaja 1837-42. Professorin arvonimi 1844. Halvautui 1863. Saavuttanut Suomen kansallisuudelle aseman. † Porvoossa 6.5.1877.

Photo: 1831: Fredrika Charlotta Tengström († 1879).

Figure 4: Person instance page of Johan Ludvig Runeberg (1804–1877) with six tabs for data analyses

biographical information in the traditional way as well as related information and analyses based on the entire material and supplementary data outside the registers. The instance pages are entered by clicking on the person in the faceted search result list. For example, Figure 4 shows the page of Johan Ludvig Runeberg (1804–1877). Instance pages are presented in a table format where the left column lists the characteristics associated with the person, such as university enrollment time, vocation, and student nation. On the right are the values, such as the time of enrollment on 10 February 1822 and the vocations of “Lecturer at Porvoo High School”, “Docent” and “Author”, as well as links to further information. People are linked to Wikidata/Wikipedia and BiographySampo. Wikidata links have been used to retrieve people’s Wikipedia pages and possible photos of people from Wikimedia Commons. Correspondingly,

links to the National Biography of Finland¹⁴ and BiographySampo¹⁵ have been retrieved from the BiographySampo LOD service. Links to the dissertation publications in the Doria¹⁶ archive were already available in the source material. The various Sampo projects with their underlying ontologies and data are gradually expanding into a kind national Cloud of LOD services, a kind of “SampoSampo” that is also connected to the international Linked Open Data (LOD) cloud¹⁷.

Analyzing and Visualizing an Individual Person One of the innovations of the Sampo model and portals is to offer the user, in addition to intelligent search and browsing functions, data analytical tools and visualizations for more detailed content exploration and knowledge discovery [13]. Tools can be selected from the tabs at the top of the search views (see Figure 2), applying the tool to the set of search results found by facet selections in the current view. In the same vein, there are also tabs for data analytical tools on the different instance pages for examining particular individuals (instances). For example, in addition to the default table view shown in Figure 4 (TABLE tab), the user can choose to explore a person’s relationships, academic relationships with other students, relationships with other people in the same organizations, or a network of events related to the person.

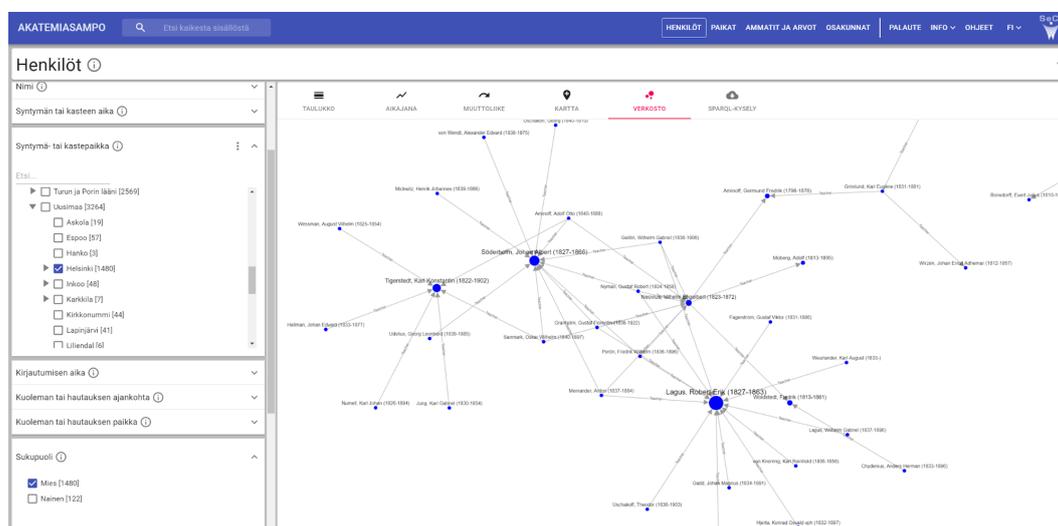


Figure 5: Academic teacher-student networks shown on the NETWORK tab

The FAMILY RELATIONS tab on the page of Johan Ludvig Runeberg (1804–1877)¹⁸ shows the network of his close relatives to a limited depth. Its data has been extracted from people mentioned in various contexts in the ACADEMYSAMPO, and kinship has also been enriched by reasoning. The network shows, for example, that the wife of Runeberg’s son, sculptor Walter

¹⁴<https://kansallisbiografia.fi/english/national-biography>

¹⁵<https://biografiasampo.fi>

¹⁶<https://www.doria.fi/>

¹⁷Linked Open Data Cloud: <https://www.lod-cloud.net>

¹⁸<https://akatemiasampo.fi/en/people/page/p13687/familyNetwork>

students (597) 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 pages. The MAP tab (Figure 9) 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. 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 (Figure 5).

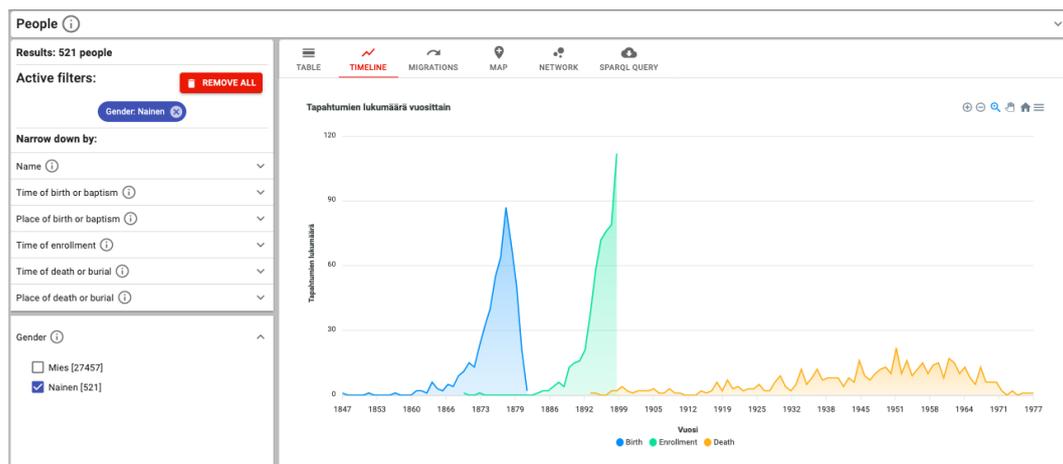


Figure 7: The annual births, enrollments, and deaths of female students on the TIMELINE tab

(2) Places Perspective The Places view of the portal²² offers a similar faceted search as in the People view, but now targeting historical places. You can search for places using the hierarchical facet or text search and see the search results as a table. In Figure 9, the user has searched ACADEMYSAMPO’s more than 3000 historical sites for those with the word “university” in their name. The result contains 14 universities from Finland, Sweden, Estonia, and the rest of Europe with photos from Wikidata, such as the seven German universities mentioned in the registers. For Finnish place names, the location material, e.g., coordinate location, alternative labels, and hierarchy of places, has been extracted from the National Land Survey of Finland’s PNR database (Place Name Register)²³ and from the YSO Places ontology of the National Library’s Finto.fi²⁴ ontology service. The most important source for foreign places has been Wikidata providing also possible photographs and coats of arms. However, especially the earlier material contains many references to towns and villages in Sweden, which were geocoded using data from the international GeoNames database²⁵.

²²Places Perspective: <https://akatemiasampo.fi/en/places/faceted-search>

²³Available as LOD at <https://www.ldf.fi/dataset/pnr/index.html>

²⁴YSO Places: <https://finto.fi/yso-paikat/en/>

²⁵GeoNames: <https://www.geonames.org>

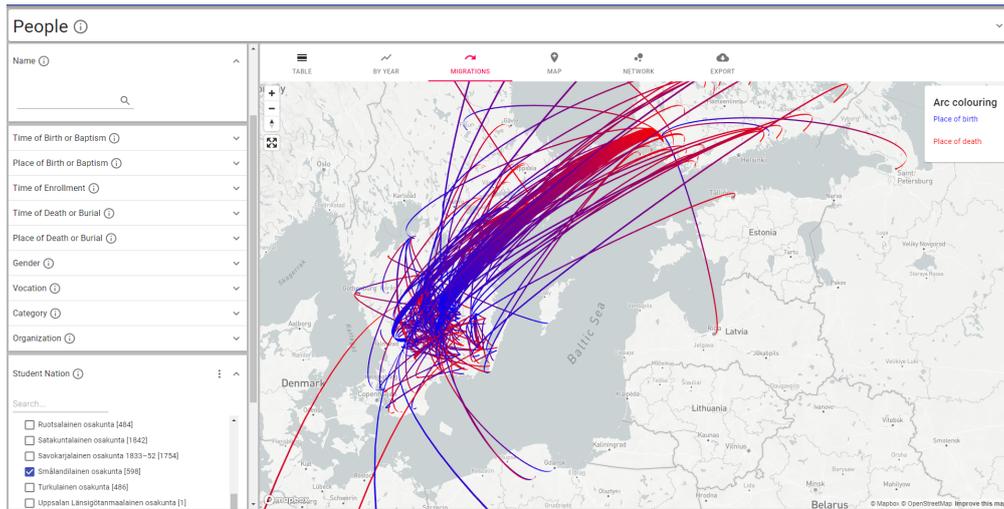


Figure 8: The use of ACADEMYSAMPO for prosopographical research. Curricula vitae of members of the Student Nation of Småland with the places of birth (blue end of the arch) and death (red end)

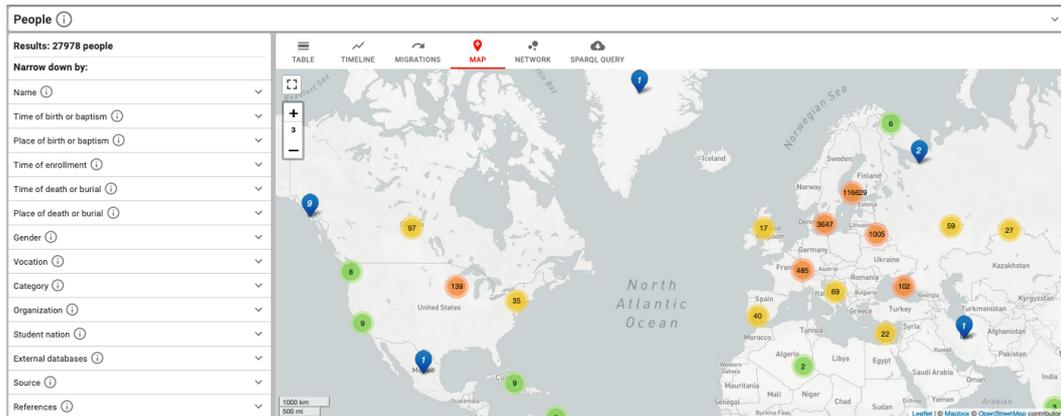


Figure 9: The MAP tab visualizes the lifetime places mentioned in about 175 000 events

The search result can also be visualized on a separate MAP tab²⁶ (Figure 9). Clicking on a place marker on the map opens a pop-up window with a list of the people whose events, such as death or career-related events, are known to have occurred at that location. A link in the pop-up window leads to the person's instance page for a more detailed investigation of the event. Figure 10 depicts the MAP tab zoomed to downtown Helsinki where you can find e.g. Helsinki Normal Lyceum²⁷. The placement of Norssi on the map is an example of the possibilities of linked data: only textual mentions of the Normal Lyceum are found in the registers, but the

²⁶MAP tab: <https://akatemiasampo.fi/en/places/faceted-search/map>

²⁷Wikidata resource for Helsinki Normal Lyceum: <https://www.wikidata.org/wiki/Q3269135>

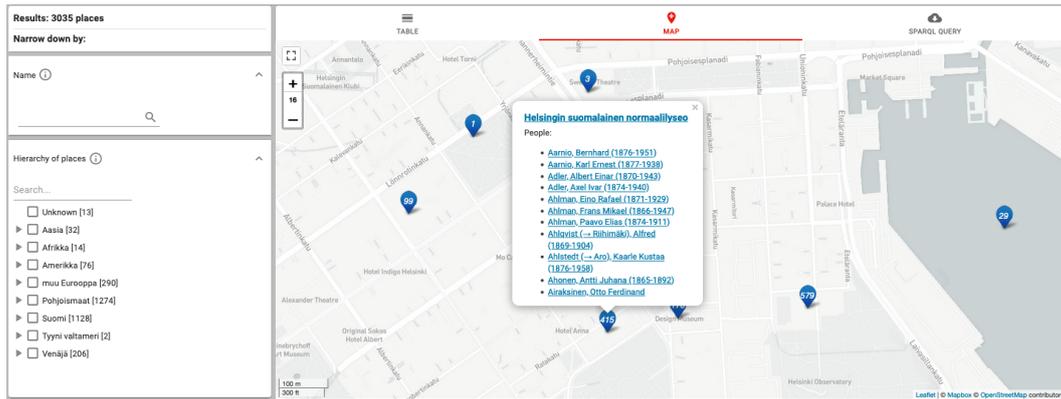


Figure 10: The MAP tab in Places view shows the places and the people associated with them through events. This view shows 415 students and other people connected to the Helsinki Normal Lyceum.

open-licensed geo-information in Wikidata can also be used in ACADEMYSAMPO. Clicking on the marker opens a pop-up window listing all 415 students and other school-related people found in the ACADEMYSAMPO, such as Ivar Edvard Wilskman (1854–1932), a school gymnastics teacher and professor known as the father of Finnish sports. Further information about the students of Norssi can be found in the online service “Norssi Alumni on the Semantic Web” [14], one of the biographical Sampo systems preceding ACADEMYSAMPO.

(3) Vocations Perspective The Vocations view²⁸ allows to search for people and groups by vocations, as well as places related to the person. The classification of vocations is based on the AMMO ontology of historical vocations [15], which is linked to, e.g., to the international HISCO classification²⁹. The AMMO ontology offers opportunities to study, for example, the social status of students or the inheritance of vocations across generations.

(4) Students Nations Perspective Student Nations have formed an important part of the student lives at the universities, bringing together students from the same area and creating links between the students and the university administration. The institution of student nations was established at the Royal Academy of Turku in 1643. Some of the current student nations of the University of Helsinki are heirs of the original departments, but many nations have been divided or merged into new nations over time. Completely new student nations have also been established while old ones have been abolished. In the Student Nations view³⁰ you can search for nations having their own instance pages similarly as people and places. For example, members of the student nation at different times, curators, inspectors and honorary members have been gathered as links to their websites, insofar as they are mentioned in the register texts. In addition, the data includes references to Swedish student nations for example at the universities of Uppsala and Lund.

²⁸Vocations Perspective: <https://akatemiasampo.fi/en/titles/faceted-search>

²⁹HISCO classification: <https://iisg.amsterdam/en/data/data-websites/history-of-work>

³⁰Student Nations Perspective: <https://akatemiasampo.fi/en/studentNations/faceted-search>

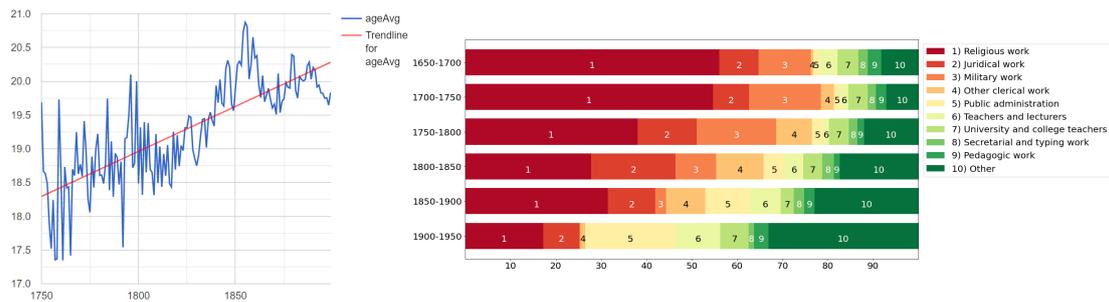


Figure 11: Left: average age of students at enrollment is raising. Right: most common ten vocational categories in different times 1650–1950; Religious work declines while Public administration raises

4. Using the SPARQL Endpoint for Data Analysis

A goal of this article is to encourage DH researchers to take advantage of the ACADEMYSAMPO LOD service using SPARQL querying. Therefore, examples of the use of YASGUI, Google Colab, or Jupyter notebooks are introduced in the following. These tools provide a simple way to create analyzes of data and share the results as functional documents for others to evaluate and use according to the principles of open science.

YASGUI provides an easy-to-use web-based browser editor for writing SPARQL queries. The answers to the queries can be examined, e.g., in tabular form, and can also be easily visualized in pre-programmed ways, such as displaying location data on a map, or generating various charts. For example, the chart on the left in Figure 11 depicts the average age of students with respect to the year of enrollment. The trendline shows how the age at enrollment has evolved during the years shown in the chart, with an obvious increase between about 1825 and 1850. The query itself requires just over ten lines of SPARQL query language. This result has been visualized using a ready-made chart tool of YASGUI.

Figure 12 shows visualization of the geographical data using the YASGUI editor. The markers on the map are colored according to the student nation to which the majority of student born there belonged to. It can be seen that the base areas of the Student Nations are formed quite clearly. One can see, for example, that the Baltic countries and Russia have been the base area of the student nation of Vyborg. Implementing this visualization has required only a slightly more complex query to retrieve the data in appropriate format.

For more complex analysis or more customized options, the results of a SPARQL query can be analyzed using libraries in different programming languages. Google Colab provides an easy way to write and run Python code online as Jupyter notebooks using a web browser, edit them collaboratively, and share results easily and visually. A document can consist of explanatory text snippets, code snippets that can be interpreted, and visualizations produced using Python code and libraries. The chart on the right in Figure 11 shows the percentages of the most common vocational categories during the years 1650–1950 in intervals of 50 years. By looking at the figure one can observe how religious and juridical work have been the most significant ones during the early centuries while when approaching the 20th century vocations related to e.g. administration and education gain more importance. This visualization was created in Google

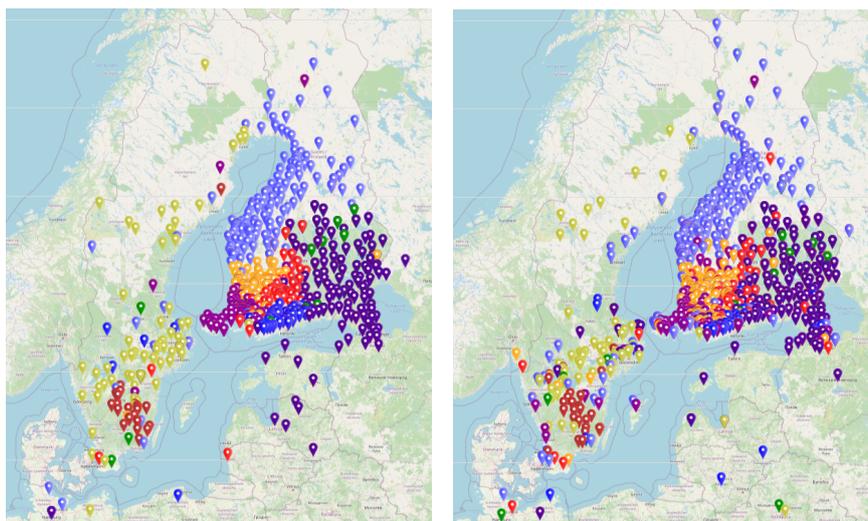


Figure 12: Map visualizations using spatial data in the YASGUI editor. The places of birth (on the left) and death (on the right) are colored based on the student nation having the most students at that place. The visualization can be tested at https://api.triptydb.com/s/xcJe_Hj0n

Colab using Python library Matplotlib³¹, after downloading relevant data with a SPARQL query.

The data service on the Linked Data Finland platform³² provides more information and documentation about the data publication and the SPARQL endpoint. The publication follows Tim Berners-Lee's five-star model³³, but the seven-star model proposed in the LDF platform gives a sixth star because the data release also includes a description of the data model to facilitate data reuse. The seventh star would require data validation, which has not been done systematically at this stage.

5. Discussion

Related Work Representing and analyzing biographical data is 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 [16], and the proceedings of BD2017 contain more similar works [17]. In [18], analytic visualizations were created based on U.S. Legislator registry data. The idea of biographical network analysis is related to the Six Degrees of Francis Bacon system³⁴ [19, 20] that utilizes data of the Oxford Dictionary of National Biography. However, in our case, faceted search can be used for filtering and studying target groups.

Work on ACADEMYSAMPO is continuation to our earlier biographical LOD systems on Norssit Alumni register [14], the U.S. Congress Prosopographer [21], and BiographySampo [10]. Our

³¹<https://matplotlib.org>

³²ACADEMYSAMPO data service: <http://www.ldf.fi/dataset/yoma>

³³Five Star Model for publishing Linked Data:

<https://www.w3.org/community/webize/2014/01/17/what-is-5-star-linked-data/>

³⁴<http://www.sixdegreesoffrancisbacon.com>

earlier articles provide examples using Colab for analyses with the data of BiographySampo [22] and for Members of Parliament in Finland in the project ParliamentSampo [23]. Extracting Linked Data from texts has been studied in several works [24]. In [25] language technology was used for extracting entities from biographies and in [26] from news.

Data Literacy Needed The entities, concepts and relationships identified from the register texts form the basis for ACADEMYSAMPO's links, search functionalities, data analysis and visualizations. The structured metadata data of a system like ACADEMYSAMPO is largely automatically generated and requires a new kind of data literacy to use it [27]. The structures and links highlighted by the system are based on the original texts, which may be incomplete or incorrect in some respects. In addition, the algorithms used may not be able to identify all the desired expressions from the text, and errors may occur in the identification of the entities in the data and in the identification of their meanings. Identifiable concepts can also be incompatible with each other (e.g., vocational titles from different eras) and their meaning can change over time (e.g., historical places and regions). Systems based on linked ontological data strongly highlight data inconsistencies, errors, and omissions in the user interface.

For example, there are both discontinued and still operating student nations in the database, and the place ontology includes areas which were previously part of Finland and Sweden but later annexed by the Soviet Union. Such data presentation challenges stem not so much from the methods of linked data as from the ontological complexity of the real world being described and the shortcomings and inaccuracies associated with historical register data, but defining and using ontologies in search, browsing, and data analysis reveals data structures. In traditional search systems, problems are somehow swept under the rug in textual data and human interpretation when reading data. The aim of ACADEMYSAMPO is to facilitate the researcher's work in reviewing and researching register material by automatically extracting interesting references, links and visualizations whenever technically possible, an example of the "distant reading" of digital humanities [28]. Using semantically linked, rich data easily creates a misconception that the data and links would be complete and the gaps in the data would be flawed. It must be recalled that the knowledge extracted by a system such as ACADEMYSAMPO is, of course, based only on the available data. For example, for people who do not have their own article in the register, such as most of the students' wives and relatives or characters like James Cook, no information other than mentions in the descriptions of the register people is available.

Acknowledgements Yrjö Kotivuori and Veli-Matti Autio authored the original data publications used in our work. Our work is related to the EU project InTaVia: In/Tangible European Heritage³⁵. CSC – IT Center for Science has provided computational resources for the work.

References

- [1] 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, CEUR Workshop Proceedings, Vol. 2612, 2020, pp. 284–292. URL: <http://ceur-ws.org/Vol-2612/short14.pdf>.

³⁵<https://intavia.eu/>

- [2] E. Hyvönen, P. Leskinen, H. Rantala, E. Ikkala, J. Tuominen, Akatemiasampo-portaali ja -datapalvelu henkilöiden ja henkilöryhmien historialliseen tutkimukseen (AcademySampo portal and data service for biographical and prosopographical research), *Informaatio-tutkimus* 40 (2021) 28–56. URL: <https://journal.fi/inf/article/view/102656>.
- [3] 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: *ESWC 2014: The Semantic Web: ESWC 2014 Satellite Events*, Springer-Verlag, 2014, pp. 226–230. doi:10.1007/978-3-319-11955-7_24.
- [4] K. Verboven, M. Carlier, J. Dumolyn, A short manual to the art of prosopography, in: *Prosopography approaches and applications. A handbook*, Unit for Prosopographical Research (Linacre College), 2007, pp. 35–70. doi:1854/8212.
- [5] L. Rietveld, R. Hoekstra, The YASGUI family of SPARQL clients, *Semantic Web – Interoperability, Usability, Applicability* 8 (2017) 373–383. doi:10.3233/SW-150197.
- [6] E. Hyvönen, Digital humanities on the Semantic Web: Sampo model and portal series, *Semantic Web – Interoperability, Usability, Applicability* (2022). URL: <http://www.semantic-web-journal.net/content/digital-humanities-semantic-web-sampo-model-and-portal-series>, accepted.
- [7] 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.
- [8] E. Hyvönen, Linked open data infrastructure for digital humanities in Finland, in: *DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, CEUR Workshop Proceedings, vol. 2612, 2020, pp. 254–259. URL: <http://ceur-ws.org/Vol-2612/short10.pdf>.
- [9] T. Heath, C. Bizer, *Linked Data: Evolving the Web into a Global Data Space* (1st edition), *Synthesis Lectures on the Semantic Web: Theory and Technology*, Morgan & Claypool, 2011. URL: <http://linkeddatabook.com/editions/1.0/>.
- [10] 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: *Proceedings of the 16th Extended Semantic Web Conference*, Springer-Verlag, 2019, pp. 574–589.
- [11] P. Leskinen, E. Hyvönen, Extracting Genealogical Networks of Linked Data from Biographical Texts, in: *The Semantic Web: ESWC 2019 Satellite Events*, Springer, 2019, pp. 121–125.
- [12] P. Leskinen, E. Hyvönen, Reconciling and Using Historical Person Registers as Linked Open Data in the AcademySampo Knowledge Graph, in: *Proceedings of the 20th International Semantic Web Conference (ISWC 2021)*, Springer-Verlag, 2021, pp. 714–730. doi:10.1007/978-3-030-21348-0_37.
- [13] E. Hyvönen, Using the semantic web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery, *Semantic Web* 11 (2020) 187–193. doi:10.3233/SW-190386.
- [14] 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: *Language, Technology and Knowledge*, Springer-Verlag, 2017, pp. 113–119.

- [15] M. Koho, L. Gasbarra, J. Tuominen, H. Rantala, I. Jokipii, E. Hyvönen, AMMO Ontology of Finnish Historical Occupations, in: Proceedings of the The First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH'19), volume 2375, CEUR Workshop Proceedings, 2019, pp. 91–96. URL: <http://ceur-ws.org/Vol-2375/>, vol 2375.
- [16] S. ter Braake, R. S. Anstke Fokkens, T. Declerck, E. Wandl-Vogt (Eds.), BD2015, Biographical Data in a Digital World 2015, CEUR Workshop Proceedings, Vol-1399, 2015. URL: <http://ceur-ws.org/Vol-1399/>.
- [17] A. Fokkens, S. ter Braake, R. Sluijter, P. Arthur, E. Wandl-Vogt (Eds.), BD2017 Biographical Data in a Digital World 2015, CEUR Workshop Proceedings, Vol-2119, 2017. URL: <http://ceur-ws.org/Vol-2119/>.
- [18] R. Larson, Bringing lives to light: Biography in context, 2010. Final Project Report, University of Berkeley, http://metadata.berkeley.edu/Biography_Final_Report.pdf.
- [19] 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).
- [20] A. Langmead, J. Otis, C. Warren, S. Weingart, L. Zilinski, Towards interoperable network ontologies for the digital humanities, *Int. J. of Humanities and Arts Computing* 10 (2016) 22–35.
- [21] G. Miyakita, P. Leskinen, E. Hyvönen, Using linked data for prosopographical research of historical persons: Case U.S. Congress Legislators, in: 7th International Conference, EuroMed 2018, Proc., Part II, Springer-Verlag, 2018, pp. 150–162.
- [22] 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* (2021). Accepted.
- [23] P. Leskinen, E. Hyvönen, J. Tuominen, Members of parliament in finland knowledge graph and its linked open data service, in: Further with Knowledge Graphs. Proceedings of the 17th International Conference on Semantic Systems, 6-9 September 2021, Amsterdam, The Netherlands, IOS Press, 2021, pp. 255–269. URL: <https://ebooks.iospress.nl/volumearticle/57420>. doi:10.3233/SSW210049.
- [24] 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.
- [25] A. Fokkens, S. ter Braake, N. Ockeloën, P. Vossen, S. Legêne, G. Schreiber, V. de Boer, BiographyNet: Extracting Relations Between People and Events, in: *Europa baut auf Biographien*, New Academic Press, Wien, 2017, pp. 193–224.
- [26] 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 World Wide Web* 37 (2016) 132–151.
- [27] T. Koltay, Data literacy for researchers and data librarians, *Journal of Librarianship and Information Science* 49 (2017) 3–14. doi:10.1177/0961000615616450.
- [28] F. Moretti, *Distant Reading*, Verso Books, 2013.