Reassembling and Enriching the Life Stories in Printed Biographical Registers: Norssi High School Alumni on the Semantic Web

Eero Hyvönen, Petri Leskinen, Erkki Heino, Jouni Tuominen, and Laura Sirola

Semantic Computing Research Group (SeCo), Aalto University and
HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
http://seco.cs.aalto.fi, http://heldig.fi
firstname.lastname@aalto.fi

Abstract. This paper presents the idea to enrich printed biographical person registers with linked data related to events that took place after the register was published. By transforming printed historical documents into structured data, semantic search to written texts can be provided for the reader. Even more importantly, life stories of historical persons can be extended based on data linking by extracting semantic structures from printed texts, and by combining this data with external datasets and data services. Such linking provides an enriched context for prosopographical research on people in the register, as well as an enhanced reading experience for anyone interested in reading the biographies. As a concrete case study, a register 1867–1992 of over 10 000 alumni of the prominent Finnish high school “Norssi” was transformed into RDF, was enriched by data linking, was published as a linked data service, and is provided to end users via a faceted search engine and browser for studying lives of historical persons and for prosopographical research.

1 Biographical Registers

Schools, professional guilds, scientific societies, and other person organizations regularly publish biographical registers of their members. Such registers provide a valuable source of information on personal data of groups of people. At the same time, social cohesion and self-esteem of people sharing e.g. common history, interests, or other aspects of life can be enhanced. To name a few examples in Finland, the government has regularly published the “State Calendar” (Suomen Valtiokalenteri)[1] of prominent Finnish officials, the historical Student Register (Ylioppilasmatrikkeli)[2] 1640–1852 of the University of Helsinki contains data about 18 000 early academic persons in Finland, and there is a register of 73 100 engineers and architects in Finland[3], maintained by the labor union TEK since 1930’s. Registers are usually created while the persons listed are still alive.

[1] https://www.valtiokalenteri.fi/

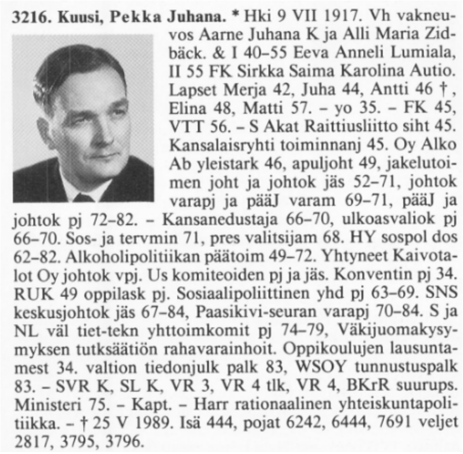
[2] http://www.helsinki.fi/ylioppilasmatrikkeli/

[3] https://fi.wikipedia.org/wiki/Tekniikan_akateemiset_ja_arkkitehdit_-matrikkeli

Such registers typically contain short biographical entries of people that belong to some group, with perhaps a photo attached. Traditionally, such registers have been published in print, making it difficult to keep the data up-to-date. When reading an old register, a recurring problem is to find out what happened to the persons after the register was published. For example, when reading one's old high school graduation register: what happened to the classmates afterwards?

This paper presents an overview of research underway, addressing the problem of transforming printed biographical registers into Linked Data, and enriching their contents using Named Entity Linking [3,2]. As a concrete case study, we consider the printed register "Norssit 1867–1992. Helsingin Norssin matrikkeli", a book of 708 pages, containing short bios of over 10 000 students and teachers of the prominent Finnish high school "Norssi", a training school of the University of Helsinki. This school celebrates its 150th anniversary in 2017, so this is a good moment to create an enriched look back at the history of its alumni.

2 Norssi Alumni on the Semantic Web



3216. Kuusi, Pekka Juhana. * Hki 9 VII 1917. Vh vakneuvos Aarne Juhana K ja Alli Maria Zidbäck. & I 40–55 Eeva Anneli Lumiala, II 55 FK Sirkka Saima Karolina Autio. Lapset Merja 42, Juha 44, Antti 46 †, Elina 48, Matti 57. – yo 35. – FK 45, VTT 56. – S Akat Raittiusliitto siht 45. Kansalaisryhti toiminnanj 45. Oy Alko Ab yleistark 46, apuljoht 49, jakelutoimen joht ja johtok jäs 52–71, johtok varapj ja pääJ varam 69–71, pääJ ja johtok pj 72–82. – Kansanedustaja 66–70, ulkoasvaliok pj 66–70. Sos- ja tervmin 71, pres valitsijam 68. HY sospol dos 62–82. Alkoholipolitiikan päätoim 49–72. Yhtyneet Kaivotalot Oy johtok vpj. Us komiteoiden pj ja jäs. Konventin pj 34. RUK 49 oppilask pj. Sosiaalipoliittinen yhd pj 63–69. SNS keskusjohtok jäs 67–84, Paasikivi-seuran varapj 70–84. S ja NL väl tiet-tekn yhttoimkomit pj 74–79, Väkijuomakysymyksen tutksäätiön rahavarainhoit. Oppikoulujen lausuntamest 34. valtion tiedonjulk palk 83, WSOY tunnustuspalk 83. – SVR K, SL K, VR 3, VR 4 tlk, VR 4, BKrR suurups. Ministeri 75. – Kapt. – Harr rationaalinen yhteiskuntapolitiikka. – † 25 V 1989. Isä 444, pojat 6242, 6444, 7691 veljet 2817, 3795, 3796.

Fig. 1. A short biographical entry in the register book Norssit 1867–1992

Extracting Structure from Text The project started by digitizing the the book at the Digitization Centre of the National Library of Finland. As a result, an OCR-version in XML of the book pages was obtained, including coordinates of detected images of persons. The data extracted was then transformed into RDF form, where each biographical entry was extracted from the OCR text. Also the photos of persons were extracted from the images of the book pages and linked with the bios. After this, a collection of regex rules and Python scripts were designed in order to 1) clean OCR errors in the data and to 2) extract various pieces of information from the short bios, such as the name of the person, birth place, hobbies, and relatives mentioned. An example of a short biograph in the book is depicted in Fig. 1. The extracted data was then uploaded into a SPARQL endpoint of the Linked Data Finland service[4] [5].

From a data linking viewpoint, the birthday and full name of the persons were known at this point, which could be used to enrich the data from several other datasets listed in Table 1. Links were created to Wikipedia, Wikidata, National Biography of Finland[5] and its Swedish complement BLF[6], BookSampo[7] Linked Data, CultureSampo[8] portal, WarSampo[9] portal, ULAN[10] authority register by The J. Paul Getty Trust, VIAF[11], and the genealogical data service Geni[12]. For entity linking to databases offering a SPARQL endpoint, the tool SPARQL ARPA[13] was used. In cases where the database provides a REST API, like Wikipedia or Geni.com, a special Python script was used. The script was used also in the case of BLF, where the data was available as a CSV formatted table.

Data Source	**Links**	**Description**
Wikipedia	496	`http://fi.wikipedia.org`
Wikidata	501	`http://www.wikidata.org`
National Biography	136	National Biography of Finland
BLF	44	Biografiskt Lexikon för Finland
BookSampo	90	Finnish fiction literature on the Semantic Web service
CultureSampo	453	LOD from museums, archives, libraries, and media
WarSampo	353	Second World War LOD service and portal
ULAN	21	Union List of Artist Names Online
VIAF	135	Virtual International Authority Files
Geni	891	Family research and family tree data

Table 1. Data sources linked to the Norssit register

For example, the RDF data corresponding to Fig. 1 is presented below (with long URIs and literal values shortened for brevity by using three periods):

```
@prefix schema: <http://schema.org/> .
@prefix registry: <http://ldf.fi/schema/person_registry/> .
```

[4] http://ldf.fi

[5] http://www.kansallisbiografia.fi/english

[6] http://www.sls.fi/sv/projekt/blf-biografiskt-lexikon-finland

[7] http://www.kirjasampo.fi

[8] http://www.kulttuurisampo.fi

[9] http://sotasampo.fi/en/

[10] http://www.getty.edu/research/tools/vocabularies/ulan/

[11] http://www.viaf.org

[12] http://www.geni.com

[13] http://seco.cs.aalto.fi/projects/dcert/

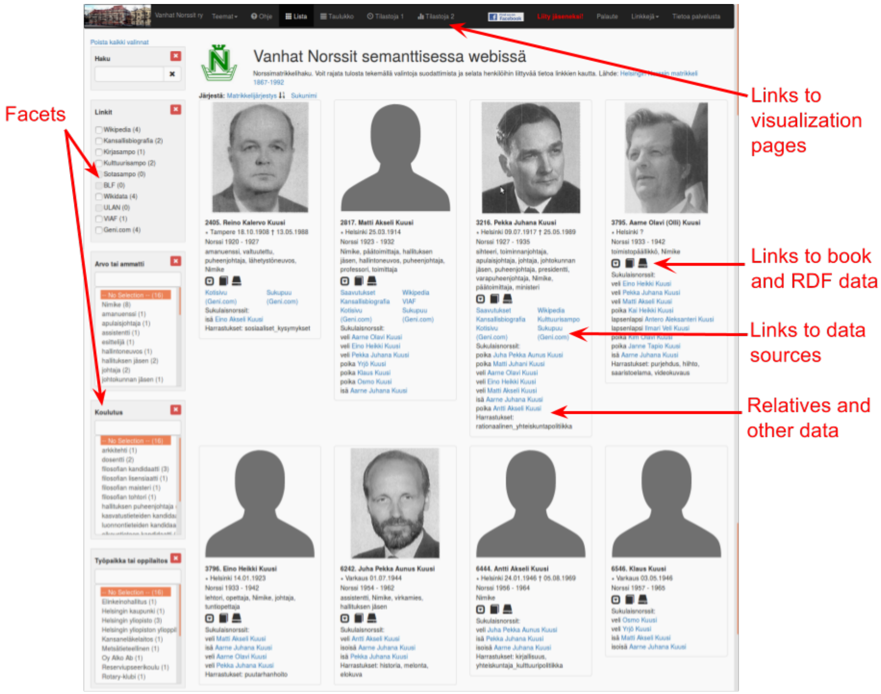
```
@prefix dct:   <http://purl.org/dc/terms/> .
@prefix hobbies: <http://ldf.fi/hobbies> .
@prefix achievement: <http://ldf.fi/norssit/achievements/> .
@prefix bioc:  <http://ldf.fi/schema/bioc/> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
@prefix imagebank: <http://static.seco.cs.aalto.fi/norssit/images/profile/> .
@prefix skos:  <http://www.w3.org/2004/02/skos/core#> .
@prefix foaf:  <http://xmlns.com/foaf/0.1/> .
@prefix norssit: <http://ldf.fi/norssit/> .
norssit:norssi_3216  a              foaf:Person ;
        achievement:works         achievement:achievement_639 ;
        bioc:has_family_relation  [ a                bioc:Brother ;
                                  bioc:inheres_in  norssit:norssi_3796 ] ;
        bioc:has_family_relation  [ a                bioc:Son ;
                                  bioc:inheres_in  norssit:norssi_7691 ] ;
        bioc:has_family_relation  [ a                bioc:Son ;
                                  bioc:inheres_in  norssit:norssi_6444 ] ;
        bioc:has_family_relation  [ a                bioc:Son ;
                                  bioc:inheres_in  norssit:norssi_6242 ] ;
        bioc:has_family_relation  [ a                bioc:Brother ;
                                  bioc:inheres_in  norssit:norssi_3795 ] ;
        bioc:has_family_relation  [ a                bioc:Brother ;
                                  bioc:inheres_in  norssit:norssi_2817 ] ;
        bioc:has_family_relation  [ a                bioc:Father ;
                                   bioc:inheres_in  norssit:norssi_444 ] ;
        norssit:genicom             <https://www.geni.com/people/...> ;
        norssit:kansallisbiografia  <http://www.kansallisbiografia...> ;
        norssit:kulsa               <http://www.seco.tkk.fi/...> ;
        norssit:kulttuurisampo      <http://www.kulttuurisampo.fi/...> ;
        norssit:wikidata            <https://www.wikidata.org/wiki/...> ;
        norssit:wikipedia           <https://fi.wikipedia.org/...> ;
        registry:birthPlace         "Helsinki"@fi ;
        registry:enrollmentYear     "1927"^^xsd:gYear ;
        registry:entryText          "3216. Kuusi, Pekka Juhana ... "@fi ;
        registry:matriculationYear  "1935"^^xsd:gYear ;
        registry:pageImageURL       <http://static.seco.cs.aalto.fi/...> ;
        registry:pageNumber         271 ;
        dct:description             "Pekka Juhana Kuusi ..." ;
        schema:birthDate            "1917-07-09"^^xsd:date ;
        schema:birthPlace           <http://ldf.fi/places/Helsinki> ;
        schema:deathDate            "1989-05-25" ;
        schema:familyName           "Kuusi"@fi ;
        schema:gender               schema:Male ;
        schema:givenName            "Pekka Juhana"@fi ;
        schema:hobby                <http://ldf.fi/hobbies/...> ;
        schema:image                imagebank:3216.png .
```

Application Online Based on the RDF data, a faceted search and browsing application[14] depicted in Fig. 2 was created using the SPARQL Faceter tool [6]. On the left, the first column contains the following facets: 1) Text search. 2) Links to the data sources listed in Table 1. 3) Family name. 4) Place of birth. 5) Year of enrollment. 6) Year of graduation. 7) Hobbies. Each row presents a person, and the columns contain the data related to the facets. The last column depicts the original text from the register entry. By clicking on it, the page in the book from which the text comes from is shown. Especially interesting is the facet and

[14] http://www.norssit.fi/semweb

column for links to other data sources. For example, by selecting WarSampo or Wikipedia, classmates with a history in the WarSampo Second Word War history portal or Wikipedia page can be filtered, and corresponding homepages on these external services be found. In this way, the reading experience of an end user can be extended substantially.

Prosopograhical Research Furthermore, faceted search provides the end user with a means for filtering and studying subgroups of people in the register for prosopographical research, say persons having a Wikipedia page, born in the same area, having the same education or hobbies, etc. The upper bar of the application contains link buttons to two separate pages of visualizations that include, e.g., pie charts and histograms, based on Google graphics. By making filtering selections on facets as in Fig. 2, the graphics are automatically updated accordingly. For example, a pie chart there depicts the distribution of the higher education degrees of the filtered alumni subgroup, a multi-bar histogram visualizes most common professions of the filtered persons as time goes by, and yet another graph shows the popularity of different universities and colleges chosen by the alumni after the high school.



Vanhat Norssit semanttisessa webissä

Norssimatrikkelihaku. Voit rajata tulosta tekemällä valintoja suodattimista ja selata henkilöihin liittyvää tietoa linkkien kautta. Lähde: Helsingin Norssin matrikkeli 1867-1992

Järjestä: Matrikkelijärjestys Sukunimi

2405. Reino Kalervo Kuusi
⁎ Tampere 18.10.1908 † 13.05.1988
Norssi 1920 - 1927
amanuenssi, valtuutettu, puheenjohtaja, lähetystöneuvos, Nimike
Kotisivu (Geni.com) Sukupuu (Geni.com)
Sukulaisnorssit:
isä Eino Akseli Kuusi
Harrastukset: sosiaaliset_kysymykset

2817. Matti Akseli Kuusi
⁎ Helsinki 25.03.1914
Norssi 1923 - 1932
Nimike, päätoimittaja, hallituksen jäsen, hallintoneuvos, puheenjohtaja, professori, toimittaja
Saavutukset Wikipedia Kansallisbiografia VIAF Kotisivu (Geni.com) Sukupuu (Geni.com)
Sukulaisnorssit:
veli Aarne Olavi Kuusi
veli Eino Heikki Kuusi
veli Pekka Juhana Kuusi
poika Yrjö Kuusi
poika Klaus Kuusi
poika Osmo Kuusi
isä Aarne Juhana Kuusi

3216. Pekka Juhana Kuusi
⁎ Helsinki 09.07.1917 † 25.05.1989
Norssi 1927 - 1935
sihteeri, toiminnanjohtaja, apulaisjohtaja, johtaja, johtokunnan jäsen, puheenjohtaja, presidentti, varapuheenjohtaja, Nimike, päätoimittaja, ministeri
Saavutukset Wikipedia Kansallisbiografia Kulttuurisampo Kotisivu (Geni.com) Sukupuu (Geni.com)
Sukulaisnorssit:
poika Juha Pekka Aunus Kuusi
poika Matti Juhani Kuusi
veli Aarne Olavi Kuusi
veli Eino Heikki Kuusi
veli Matti Akseli Kuusi
isä Aarne Juhana Kuusi
poika Antti Akseli Kuusi
Harrastukset: rationaalinen_yhteiskuntapolitiikka

3795. Aarne Olavi (Olli) Kuusi
⁎ Helsinki ?
Norssi 1933 - 1942
toimistopäällikkö, Nimike
Sukulaisnorssit:
veli Eino Heikki Kuusi
veli Pekka Juhana Kuusi
veli Matti Akseli Kuusi
poika Kai Heikki Kuusi
lapsenlapsi Antero Aleksanteri Kuusi
lapsenlapsi Ilmari Veli Kuusi
poika Kim Olavi Kuusi
poika Janne Tapio Kuusi
isä Aarne Juhana Kuusi
Harrastukset: purjehdus, hiihto, saaristoelama, videokuvaus

3796. Eino Heikki Kuusi
⁎ Helsinki 14.01.1923
Norssi 1933 - 1942
lehtori, opettaja, Nimike, johtaja, tuntiopettaja
Sukulaisnorssit:
veli Matti Akseli Kuusi
isä Aarne Juhana Kuusi
veli Aarne Olavi Kuusi
veli Pekka Juhana Kuusi
Harrastukset: puutarhanhoito

6242. Juha Pekka Aunus Kuusi
⁎ Varkaus 01.07.1944
Norssi 1954 - 1962
assistentti, Nimike, virkamies, hallituksen jäsen
Sukulaisnorssit:
veli Antti Akseli Kuusi
isoisä Aarne Juhana Kuusi
isä Pekka Juhana Kuusi
Harrastukset: historia, melonta, elokuva

6444. Antti Akseli Kuusi
⁎ Helsinki 24.01.1946 † 05.08.1969
Norssi 1956 - 1964
Nimike
Sukulaisnorssit:
veli Juha Pekka Aunus Kuusi
isä Pekka Juhana Kuusi
isoisä Aarne Juhana Kuusi
Harrastukset: kirjallisuus, yhteiskuntaja_kulttuuripolitiikka

6546. Klaus Kuusi
⁎ Varkaus 03.05.1946
Norssi 1957 - 1965
Sukulaisnorssit:
veli Osmo Kuusi
veli Yrjö Kuusi
isä Matti Akseli Kuusi
isoisä Aarne Juhana Kuusi

Fig. 2. Faceted search for short biographies in the alumni register Norssit 1867–1992

3 Related Work and Discussion

Previous works of applying Linked Data technologies to biographical data include, e.g., [7], Biography.net[15] [8], and the Semantic National Biography of Finland [4]. The conference proceedings [1] includes several papers on bringing biographical data online, on analyzing biographies with computational methods, on group portraits and networks, and on visualizations. Complementing these works, the study of this paper focuses on extracting structure from printed biographical registers. Our work also emphasizes the idea of enriching the texts with external links to other biographical datasets, and on faceted search and browsing of biographical data for prosopographical studies. Our work continues, e.g., on developing new models of biographical data for prosopographical research, and on finalizing and evaluating the data linking process (precision and recall), and the demonstrator.

Our work is part of the Severi project[16], funded mainly by Tekes. Thanks to Vanhat Norssit for funding the digitization of the register and opening the data.

References

1. ter Braake, S., Anstke Fokkens, R.S., Declerck, T., Wandl-Vogt, E. (eds.): BD2015 Biographical Data in a Digital World 2015. CEUR Workshop Proceedings (2015), http://ceur-ws.org/Vol-1272/
2. Bunescu, R.C., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: EACL. vol. 6, pp. 9–16 (2006)
3. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R.: Evaluating entity linking with Wikipedia. Artificial Intelligence 194, 130–150 (Jan 2013)
4. Hyvönen, E., Alonen, M., Ikkala, E., Mäkelä, E.: Life stories as event-based linked data: Case Semantic National Biography. In: Proc. of ISWC 2014 Posters & Demonstrations Track. CEUR Workshop Proc. (2014), http://ceur-ws.org/Vol-1272/
5. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers. pp. 226–230. Springer–Verlag (May 2014)
6. Koho, M., Heino, E., Hyvönen, E.: SPARQL Faceter—Client-side Faceted Search Based on SPARQL. In: Troncy, R., Verborgh, R., Nixon, L., Kurz, T., Schlegel, K., Vander Sande, M. (eds.) Joint Proc. of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop. No. 1615, CEUR Workshop Proceedings (2016), http://ceur-ws.org/Vol-1615/semdevPaper5.pdf
7. Larson, R.: Bringing lives to light: Biography in context (2010), http://metadata.berkeley.edu/Biography_Final_Report.pdf, Final Project Report, U. of Berkeley
8. Ockeloen, N., Fokkens, A., ter Braake, S., Vossen, P., De Boer, V., Schreiber, G., Legêne, S.: BiographyNet: Managing provenance at multiple levels and from different perspectives. In: Proceedings of the 3rd International Conference on Linked Science - Volume 1116. pp. 59–71. LISC'13, CEUR-WS.org, Aachen, Germany (2013), http://dl.acm.org/citation.cfm?id=2874585.2874592

[15] http://www.biographynet.nl

[16] http://seco.cs.aalto.fi/projects/severi