

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

Eero Hyvönen^{1,2}(✉), Petri Leskinen¹, Erkki Heino^{1,2}, Jouni Tuominen^{1,2},
and Laura Sirola¹

¹ Semantic Computing Research Group (SeCo), Aalto University, Espoo, Finland
{eero.hyvonen,petri.leskinen,erkki.heino,jouni.tuominen,
laura.sirola}@aalto.fi

² HELDIG – Helsinki Centre for Digital Humanities,
University of Helsinki, Helsinki, Finland
<http://seco.cs.aalto.fi>
<http://heldig.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 By transforming printed historical documents into structured data, semantic search to written texts can be provided for the reader.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

¹ <https://www.valtiokalenteri.fi/>.

² <http://www.helsinki.fi/ylioppilasmatrikkeli/>.

about 18 000 early academic persons in Finland, and there is a register of 73 100 engineers and architects in Finland³, maintained by the labor union TEK since 1930's. Registers are usually created while the persons listed are still alive.

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 [2, 3]. 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

Extracting Structure from Text. The project started by digitizing 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 biography in the book is depicted in Fig. 1. The extracted data was then uploaded into a SPARQL endpoint of the Linked Data Finland service⁴ [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 Finland⁵ and its Swedish complement BLF⁶, BookSampo⁷ Linked Data, CultureSampo⁸ portal, WarSampo⁹ portal, ULAN¹⁰ authority register by

³ https://fi.wikipedia.org/wiki/Tekniikan_akateemiset_ja_arkkitehdit_-_matrikkeli.

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/>.



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

Table 1. Data sources linked to the Norssit register

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

The J. Paul Getty Trust, VIAF¹¹, and the genealogical data service Geni¹². For entity linking to databases offering a SPARQL endpoint, the tool SPARQL ARPA¹³ was used. In cases where the database provides a REST API, like Wikipedia or [Geni.com](http://www.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.

¹¹ <http://www.viaf.org>.

¹² <http://www.geni.com>.

¹³ <http://seco.cs.aalto.fi/projects/dcert/>.

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/> .
@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 application¹⁴ 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.

¹⁴ <http://www.norssit.fi/semweb>.

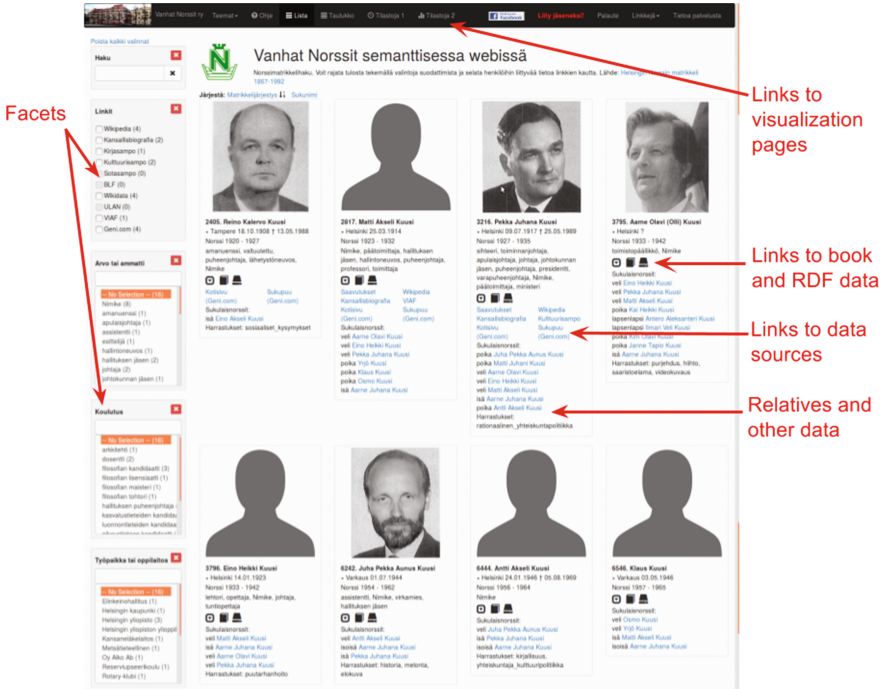


Fig. 2. Faceted search for short biographies in the alumni register Norssit 1867–1992

(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 box in the rows presents a person, containing the data related to the facets and the biography. There are also links to the original text, the RDF data, and the book page of 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 column for links to other data sources. For example, by selecting WarSampo or Wikipedia, classmates with a history in the WarSampo Second World 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 the end user can be extended substantially.

Prosopographical 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 charts. 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.

3 Related Work and Discussion

Previous works of applying Linked Data technologies to biographical data include, e.g., [7], Biography.net¹⁵ [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 project¹⁶, funded mainly by Tekes. Thanks to Vanhat Norssit for funding the digitization of the register and opening the data.

References

1. ter Braake, S., Fokkens, A., Sluijter, R., Declerck, T., Wandl-Vogt, E.: BD2015 Biographical Data in a Digital World 2015, CEUR Workshop Proceedings (2015). <http://ceur-ws.org/Vol-1399/>
2. Bunescu, R.C., Pasca, M.: Using Encyclopedic knowledge for named entity disambiguation. *EACL* **6**, 9–16 (2006)
3. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R.: Evaluating entity linking with Wikipedia. *Artif. Intell.* **194**, 130–150 (2013)
4. Hyvönen, E., Alonen, M., Ikkala, E., Mäkelä, E.: Life stories as event-based linked data: case semantic national biography. In: Horridge, M., Rospocher, M., van Ossening, J. (eds.) *Proceedings of ISWC 2014 Posters and Demonstrations Track*, CEUR Workshop Proceedings, pp. 1–4 (2014). http://ceur-ws.org/Vol-1272/paper_5.pdf
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: Pre-sutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) *ESWC 2014. LNCS*, vol. 8798, pp. 226–230. Springer, Cham (2014). doi:[10.1007/978-3-319-11955-7_24](https://doi.org/10.1007/978-3-319-11955-7_24)

¹⁵ <http://www.biographynet.nl>.

¹⁶ <http://seco.cs.aalto.fi/projects/severi>.

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., Sande, M.V. (eds.) Joint Proceedings of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop, CEUR Workshop Proceedings (2016). <http://ceur-ws.org/Vol-1615/semdevPaper5.pdf>
7. Larson, R.: Bringing lives to light: biography in context. Final Project Report, University of Berkeley (2010). http://metadata.berkeley.edu/Biography_Final_Report.pdf
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: Groth, P., van Erp, M., Kauppinen, T., Zhao, J., Keßler, C., Pouchard, L.C., Goble, C., Gil, Y., van Ossenbruggen, J. (eds.) Proceedings of the 3rd International Workshop on Linked Science, LISC 2013, pp. 59–71. CEUR Workshop Proceedings (2013). <http://ceur-ws.org/Vol-1116/paper7.pdf>