Proceedings of Digital Humanities in Nordic Countries (DHN 2020), Riga, CEUR Workshop Proceedings, March, 2020. Accepted.

Linked Open Data Service about Historical Finnish Academic People in 1640–1899

Petri Leskinen[1][0000−0003−2327−6942] and
Eero Hyvönen[1,2][0000−0003−1695−5840]

[1] Aalto University, Semantic Computing Research Group (SeCo), Finland, and
[2] University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG), Finland
http://seco.cs.aalto.fi/projects/yo-matrikkelit, http://heldig.fi, first.last@aalto.fi

Abstract. The Finnish registries "Ylioppilasmatrikkeli" 1640–1852 and 1853–1899 contain detailed biographical data about virtually every academic person in Finland during the time period. This paper presents first results on transforming these registries into a Linked Open Data service using the FAIR principles. The data is based on the student registries of the University of Helsinki, formerly the Royal Academy of Turku, that have been digitized, transliterated, and enriched with additional data about the people from various other registries. Our goal is to transform this largely textual data into Linked Open Data using named entity recognition and linking techniques, and to enrich the data further based on links to internal and external data sources and by reasoning new associations in the data. The data will be published as a Linked Open Data service on top of which tools for searching, browsing, and analyzing the data in biographical and prosopographical research are provided.

1 Biographical Dictionaries on the Web

Biographical dictionaries [18] have been published traditionally as printed book series. In 2004, the Oxford Dictionary of National Biography[3] (ODNB) was published on-line, and many major biographical dictionaries started to open their editions on the Web with search engines for finding and (close) reading biographies of interest.[4]

ODNB and other early adopters of web technology started the paradigm shift in publishing and reading biographical dictionaries on the Web. Related to our work on *BiographySampo – Finnish biographies on the Semantic web* [15] we have proposed that the next paradigm shift is to publish and use biographical dictionaries as Linked Data on the Semantic Web. [16] This paper presents first results of a new case study where this idea is applied to a new dataset: the

[3] `https://www.oxforddnb.com`

[4] On-line national biographical collections include, e.g., USA's American National Biography [1], Germany's Neue Deutsche Biographie [4], France's Nouvelle Biographie générale [5], Biography Portal of the Netherlands [2], Dictionary of Swedish National Biography [3], and National Biography of Finland[5] (NBF).

Finnish registries "Ylioppilasmatrikkeli" 1640–1899[6] that contain short biographical descriptions of 28 000 students of the University of Helsinki[7], originally the Royal Academy of Turku[8] in Finland. This publication covers a significant part of the history of Finland and the Finnish university institution, since the University of Helsinki was the only university in the country during the time frame in focus.

This paper presents an overview of research underway, addressing the problem of transforming biographical registers into Linked Data, and enriching their contents using Named Entity Recognition and Linking and by reasoning. The focus of this paper is on the data transformation process; application of the data in Digital Humanities research will be reported later. We first present the source dataset and the ontology model used for representing the biographical data in a semantic, i.e., machine "understandable" way. After this the underlying knowledge graph is discussed and its publication using the Linked Data Finland platform [17]. In conclusion, related works are discussed and relations of the work in a larger setting are summarized.

2 Source Datasets

An example of a registry entry for *Anders Israel Cajander*[9] is depicted in Fig. 1. The description starts with date or year of enrollment, in this case *11.2.1830*. After that there is the full name followed by the place and time of birth. Next there is a Finnish abbreviation *Vht* meaning parents; in the example case the father is *Zachris Johan Cajander* and the mother *Gustava Karolina Neiglick*. After that there are two lists of events; the events mentioned before the em dash (—) are related studies and academic career with the University of Helsinki; for example *Ylioppilas Helsingissä 11.2.1830.* (A student in Helsinki 11.2.1830). After the em dash there is another list of events during his career; e.g. *Äyräpään tuomiokunnan tuomari 1857* (Judge at the District Court of Äyräpää 1857). A person's death is marked with the symbol † and burial with ‡; the person in example died in Wyborg on December 18th 1901.

After the life time description there is a possible field for relatives; in the example case his spouse is mentioned first *Pso: 1841 Fredrika Emelie Schildt* where *Pso* is a Finnish abbreviation for *puoliso* (spouse). There are three relatives who also have an entry in the register, e.g. two brothers *Veli: Gustaf Adolf Cajander* and *Veli: Zakarias Cajander*, and a brother-in-law *Lanko: Berndt Vilhelm Kristoffer Schildt.* The author of the 1640–1852 dataset Yrjö Kotivuori has manually added links from a person's description to those relatives also found in the register, like the three relatives in the example case. Finally, at the

[6] The registry contains two parts: the database covering the years 1640–1852 is available in Finnish and Swedish at https://ylioppilasmatrikkeli.helsinki.fi, and the registry of 1853–1899 is at https://ylioppilasmatrikkeli.helsinki.fi/1853-1899

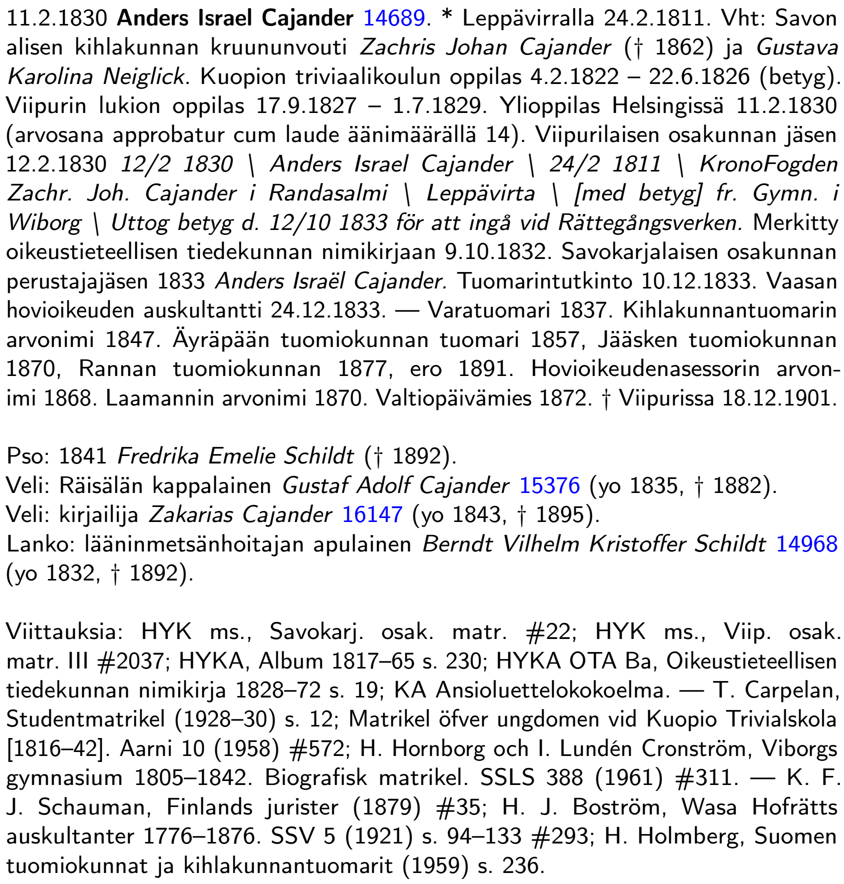
[7] https://en.wikipedia.org/wiki/University_of_Helsinki

[8] https://en.wikipedia.org/wiki/Royal_Academy_of_Turku

[9] https://ylioppilasmatrikkeli.helsinki.fi/henkilo.php?id=14689

end of the description there is a field for reference material used for collecting information about the person.

When designing the ontology model we wanted it to provide answers to possible research questions made by historians, such as: “Are there marriages between second cousins?”, “Are there families that are closely connected by marriages?”, and “What is the distance between the place of birth and the most long-term place of living and what are the mean and the standard deviation of this measure?”.



11.2.1830 **Anders Israel Cajander** 14689. * Leppävirralla 24.2.1811. Vht: Savon alisen kihlakunnan kruununvouti *Zachris Johan Cajander* († 1862) ja *Gustava Karolina Neiglick*. Kuopion triviaalikoulun oppilas 4.2.1822 – 22.6.1826 (betyg). Viipurin lukion oppilas 17.9.1827 – 1.7.1829. Ylioppilas Helsingissä 11.2.1830 (arvosana approbatur cum laude äänimäärällä 14). Viipurilaisen osakunnan jäsen 12.2.1830 *12/2 1830 \ Anders Israel Cajander \ 24/2 1811 \ KronoFogden Zachr. Joh. Cajander i Randasalmi \ Leppävirta \ [med betyg] fr. Gymn. i Wiborg \ Uttog betyg d. 12/10 1833 för att ingå vid Rättegångsverken.* Merkitty oikeustieteellisen tiedekunnan nimikirjaan 9.10.1832. Savokarjalaisen osakunnan perustajajäsen 1833 *Anders Israël Cajander.* Tuomarintutkinto 10.12.1833. Vaasan hovioikeuden auskultantti 24.12.1833. — Varatuomari 1837. Kihlakunnantuomarin arvonimi 1847. Äyräpään tuomiokunnan tuomari 1857, Jääsken tuomiokunnan 1870, Rannan tuomiokunnan 1877, ero 1891. Hovioikeudenasessorin arvonimi 1868. Laamannin arvonimi 1870. Valtiopäivämies 1872. † Viipurissa 18.12.1901.

Pso: 1841 *Fredrika Emelie Schildt* († 1892).
Veli: Räisälän kappalainen *Gustaf Adolf Cajander* 15376 (yo 1835, † 1882).
Veli: kirjailija *Zakarias Cajander* 16147 (yo 1843, † 1895).
Lanko: lääninmetsänhoitajan apulainen *Berndt Vilhelm Kristoffer Schildt* 14968 (yo 1832, † 1892).

Viittauksia: HYK ms., Savokarj. osak. matr. #22; HYK ms., Viip. osak. matr. III #2037; HYKA, Album 1817–65 s. 230; HYKA OTA Ba, Oikeustieteellisen tiedekunnan nimikirja 1828–72 s. 19; KA Ansioluettelokokoelma. — T. Carpelan, Studentmatrikel (1928–30) s. 12; Matrikel öfver ungdomen vid Kuopio Trivialskola [1816–42]. Aarni 10 (1958) #572; H. Hornborg och I. Lundén Cronström, Viborgs gymnasium 1805–1842. Biografisk matrikel. SSLS 388 (1961) #311. — K. F. J. Schauman, Finlands jurister (1879) #35; H. J. Boström, Wasa Hofrätts auskultanter 1776–1876. SSV 5 (1921) s. 94–133 #293; H. Holmberg, Suomen tuomiokunnat ja kihlakunnantuomarit (1959) s. 236.

Fig. 1. Register entry for *Anders Israel Cajander*

3 Ontology Model for Representing Biographical Data

In addition to basic data, such as people's names and dates and places of birth and death, the source data provides rich content of information like the relatives, student nation, academic degrees, career events, and sources of reference. In our case we selected the data harmonization approach and the event-centric CIDOC CRM [8] ISO standard as the ontological basis, since biographies are based on life events. Bio CRM [26] is a domain specific extension of CIDOC CRM, applicable to biographical data; it extends CIDOC CRM by introducing role-centric modeling. Bio CRM has been used in our earlier projects of Norssi High School Alumni [14,23] and BiographySampo [16,22] to model roles and occupations as well as the relationships between people.

The ontology schema is depicted in Fig. 2. The people in the register are represented as instances of the class `foaf:Person` and the mentioned relatives using `:ReferencedPerson`. The resources of actor classes are enriched with lifetime events and relationships. Events, e.g. birth, baptism, enrollment, death, and burial, are subclasses of `:Event` and enriched with linking to corresponding times, places, and titles. The source data provides two kind of binary relationships: family relations (such as *parents, children, spouses, . . .*) and domain-specific relations (such as *student, teacher, . . .*). As an example of converted RDF, the data of Fig. 1 is depicted in Fig. 3 in Turtle format.

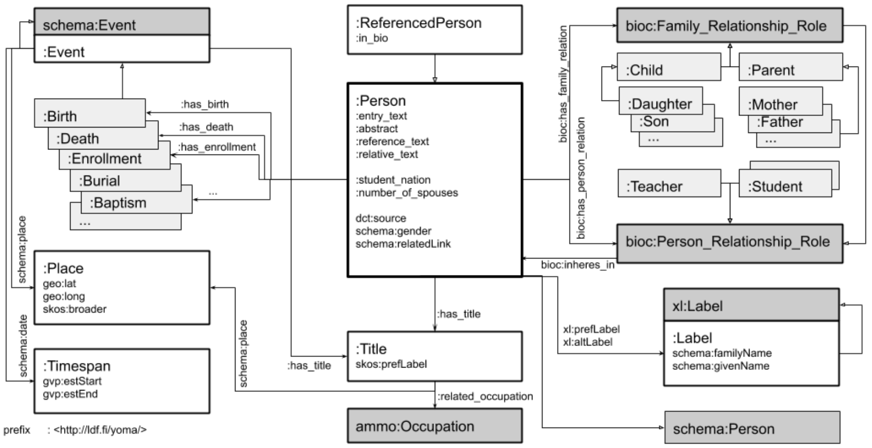


Fig. 2. Ontology schema for representing biographical data

4 Knowledge Graph of Historical Academic Persons

The extracted knowledge graph contains currently 28 000 students and 56 700 other people mentioned in the descriptions. These person resources are further

```
@prefix dct:   <http://purl.org/dc/terms/> .
@prefix foaf:  <http://xmlns.com/foaf/0.1/> .
@prefix schema: <http://schema.org/> .
@prefix skos:  <http://www.w3.org/2004/02/skos/core#> .
@prefix xl: <http://www.w3.org/2008/05/skos-xl#> .

@prefix :      <http://ldf.fi/yoma/> .
@prefix bioc:  <http://ldf.fi/schema/bioc/> .
@prefix event: <http://ldf.fi/yoma/event/> .
@prefix label: <http://ldf.fi/yoma/label/> .
@prefix rels:  <http://ldf.fi/yoma/relations/> .
@prefix titles: <http://ldf.fi/yoma/titles/> .

:p14689  a                  foaf:Person ;
        bioc:has_family_relation
                            rels:r2590153968717837790,
                            rels:r3067073318077691085,
                            ... rels:r2556529631795161483 ;
        :abstract           "<strong>Cajander, Anders Israel
                             </strong>, laamanni (yo 1830,
                             † 1901)"@fi ;
        :enrollment_text    "11.2.1830" ;
        :entry_text         "11.2.1830 <strong>Anders Israel
                            Cajander</strong> <a href= ...
                             ... († 1892).</p>\\"@fi ;
        :has_birth          event:b14689 ;
        :has_death          event:d14689 ;
        :has_enrollment     event:e1961333836730594986 ;
        :has_title          titles:v7140446880754877544 ;
        :id                 "14689" ;
        :number_of_spouses  1 ;
        :reference_text     "HYK ms., Savokarj. osak. matr. #22;
                             HYK ms., Viip. osak. matr. III #2037;
                             ... (1959) s. 236."@fi ;
        :relative_text      "<p>Veli: Räisälän kappalainen <em>
                             Gustaf Adolf Cajander</em>
                             ...  (yo 1832, † 1892).</p>"@fi ;
        :title_text         "laamanni" ;
        dct:source          :yo1640_1852 ;
        schema:gender       schema:Male ;
        schema:relatedLink
        <://ylioppilasmatrikkeli.helsinki.fi/henkilo.php?id=14689> ;
        skos:prefLabel      "Cajander, Anders Israel (1811-1901)"@fi ;
        xl:prefLabel        label:l2728541252431989123 .
```

Fig. 3. RDF/Turtle data for *Anders Israel Cajander*

enriched with 76 100 life time events and interlinked by 83 400 family and 3760 academic relations. There are 10 600 distinct occupational titles often referring to a place and an occupation, e.g., *the Bishop of Porvoo* or *Diving Commissioner who lived in Espoo.*

This information was extracted from description texts, which are structured with symbols (like † or ‡) and keywords (like *Vht* for parents or *Pso* for spouse) that help recognizing the content of each particular text field. To process the data, regular expressions, vocabularies of Finnish names, and a Python script recognizing different expressions of dates and times are used. Vocabularies of Finnish names are also used to infer a person's gender, when it is not otherwise obvious; generally, the person data is strongly male dominated, and the first female student entered the University of Helsinki in the year 1870[10].

The data set contains an ontology of more than 100 family relations. In percentage terms, most of the mentions are close relatives like *father* and *brother*, but occasionally there are, e.g., in-law-relations like *stepfather-in-law*, relations marked as potential with a question mark *son-in-law(?)*, or relations reaching over several generations like *uncle of the great great grandfather*. We extended our earlier ontology [23] to cover at least 99% of mentioned relationships. The data of the years 1640–1852 has a dense network of precise relations while in the 1853–1899 data only mentions parents and spouses; therefore, one of our future aims is to computationally extend the network to cover also the students of the 1853–1899 dataset.

The knowledge graph contains domain ontologies of 4800 place names and approximately 1000 links to our ontology of historical occupations AMMO [19]. The place ontology is the same as used in BiographySampo covering the most of Finnish towns and municipalities and also most frequently mentioned places abroad. The data will be further supplemented with ontologies of student nations and historical reference documents used as sources of information.

At this stage of work, the data has been manually validated, e.g., by making SPARQL queries or by converting the network to GraphML format [12] with the Python library NetworkX [13] and rendering it with software tools, such as Gephi [6]. Using the SPARQL queries we tested, e.g., that the years of people's birth, enrollment, and death with ages at each stage were all in a sensible range. This helped us to detect, e.g., false date or time span recognitions of our system.

A test version of the knowledge graph was published using the "7-star" Linked Data Finland model and platform (LDF.fi)[11]. LDF.fi extends Tim Berners-Lee's famous 5-star model[12] by two additional stars: the 6th start is given, if the dataset is published with the schemas it conforms to. The 7th start is given if an analysis of the quality of the data with respect to the schemas is provided, too [17]. An example of using the data service for visualizations is shown in Fig. 4, here the rich network of the family relationships of Karl Gustaf Ottelin (1792–1864).

[10] Women at the University of Helsinki (in Finnish) `http://www.helsinki.fi/yliopistonhistoria/yliopisto/nostot/naiset.htm`

[11] `http://ldf.fi`

[12] `https://5stardata.info/en/`

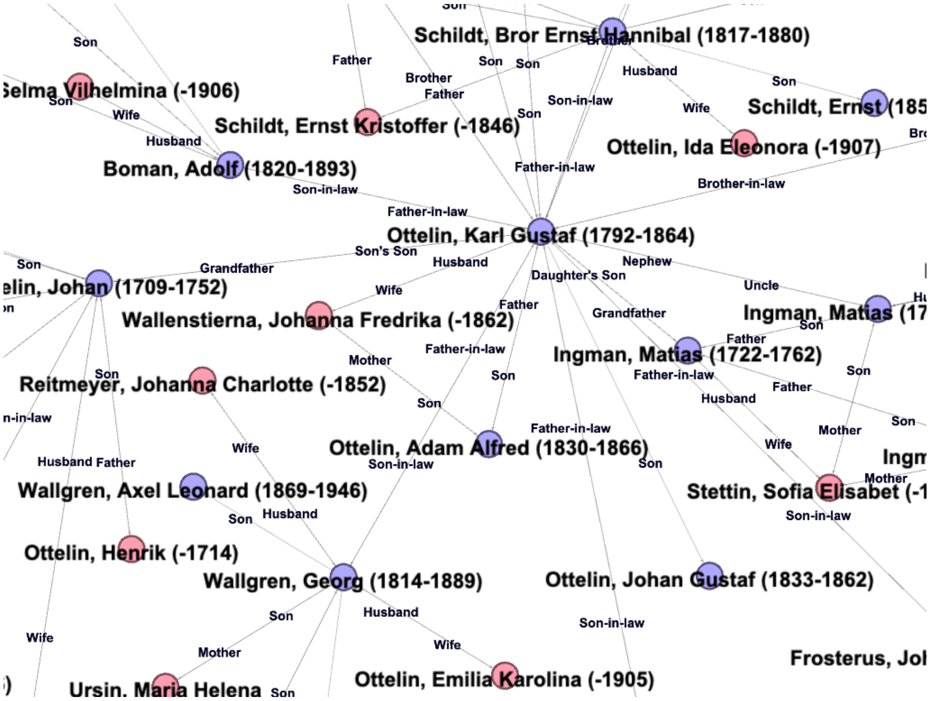


Fig. 4. Family relationships of the example person Karl Gustaf Ottelin

5 Related Work and Discussion

The Semantic Computing Research Group (SeCo) at Aalto University and University of Helsinki (HELDIG) has made earlier Linked Data publications collecting data about people in the history of Finland and beyond, including WarSampo on war history, BiographySampo, U.S. Congress Legislators [24], and Norssit High School Alumni registry [14]. The work of this paper is a continuation of these projects and further a part of a more comprehensive project of assembling an ontology of historical people in Finnish history, an important part of the emerging Linked Open Data Infrastructure for Digital Humanities in Finland initiative[13].

Representing and analyzing biographical data has grown into a new research and application field, reported, e.g., in the Biographical Data in Digital World workshops BD2015 [7], BD2017 [10], and BD2019. In [21], analytic visualizations were created based on U.S. Legislator registry data, and the Six Degrees of Francis Bacon system[14] [27,20] utilizes data of the Oxford Dictionary of National Biography. Extracting Linked Data from texts has been studied in several works, cf. e.g. [11]. In [9] language technology was applied for extracting entities and

[13] https://seco.cs.aalto.fi/projects/lodi4dh/

[14] http://www.sixdegreesoffrancisbacon.com

relations in RDF using Dutch biographies in the BiographyNet, as part of the larger NewsReader project [25].

Acknowledgements

Thanks to Yrjö Kotivuori and Veli-Matti Autio for their seminal work in creating the original Ylioppilasmatrikkeli databases used in our work, and for making the data openly available for our project to be published as Linked Open Data later on. Laura Sirola created a first RDF version of the datasets. CSC – IT Center for Science, Finland has provided computational server resources for our data service and applications.

References

1. American National Biography (2017), http://www.anb.org/aboutanb.html
2. Biography Portal of the Netherlands (2017), http://www.biografischportaal.nl/en
3. Dictionary of Swedish National Biography (2017), https://sok.riksarkivet.se/Sbl/Start.aspx?lang=en
4. Neue Deutsche Biographie (2017), http://www.ndb.badw-muenchen.de/ndb_aufgaben_e.htm
5. Nouvelle Biographie générale (2017), https://fr.wikipedia.org/wiki/Nouvelle_Biographie_generale
6. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. In: Third international AAAI conference on weblogs and social media (2009)
7. ter Braake, S., Anstke Fokkens, R.S., Declerck, T., Wandl-Vogt, E. (eds.): BD2015, Biographical Data in a Digital World 2015. CEUR Workshop Proceedings, Vol-1399 (2015), http://ceur-ws.org/Vol-1272/
8. Doerr, M.: The CIDOC CRM—an ontological approach to semantic interoperability of metadata. AI Magazine **24**(3), 75–92 (2003), https://doi.org/10.1609/aimag.v24i3.1720
9. Fokkens, A., ter Braake, S., Ockeloen, N., Vossen, P., Legêne, S., Schreiber, G., de Boer, V.: BiographyNet: Extracting Relations Between People and Events. In: Europa baut auf Biographien. pp. 193–224. New Academic Press, Wien (2017)
10. Fokkens, A., ter Braake, S., Sluijter, R., Arthur, P., Wandl-Vogt, E. (eds.): BD2017 Biographical Data in a Digital World 2015. CEUR Workshop Proceedings, Vol-1399 (2017), http://ceur-ws.org/Vol-2119/
11. Gangemi, A., Presutti, V., Recupero, D.R., Nuzzolese, A.G., Draicchio, F., Mongiovì, M.: Semantic web machine reading with FRED. Semantic Web Journal **8**, 873–893 (2017)
12. GraphML Team: The GraphML File Format, http://graphml.graphdrawing.org/ accessed: 30 September 2019
13. Hagberg, A., Swart, P., S Chult, D.: Exploring network structure, dynamics, and function using networkx. Tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States) (2008)
14. Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., Sirola, L.: Reassembling and enriching the life stories in printed biographical registers: Norssi high school alumni on the semantic web. In: Language, Technology and Knowledge. pp. 113–119. Springer–Verlag (2017)

15. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., Keravuori, K.: BiographySampo – publishing and enriching biographies on the semantic web for digital humanities research. In: Proceedings of the 16th Extended Semantic Web Conference (ESWC 2019). Springer–Verlag (2019)
16. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., Keravuori, K.: Linked data – a paradigm change for publishing and using biography collections on the semantic web. In: Proceedings of the Third Conference on Biographical Data in a Digital World (BD 2019) (September 2019)
17. 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: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) The Semantic Web: ESWC 2014 Satellite Events. ESWC 2014. pp. 226–230. Springer-Verlag (May 2014), https://doi.org/10.1007/978-3-319-11955-7_24
18. Keith, T.: Changing conceptions of National Biography. Cambridge University Press (2004)
19. Koho, M., Gasbarra, L., Tuominen, J., Rantala, H., Jokipii, I., Hyvönen, E.: AMMO Ontology of Finnish Historical Occupations. In: Proceedings of the The First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH'19). vol. 2375, pp. 91–96. CEUR Workshop Proceedings (June 2019), http://ceur-ws.org/Vol-2375/, vol 2375
20. Langmead, A., Otis, J., Warren, C., Weingart, S., Zilinski, L.: Towards interoperable network ontologies for the digital humanities. Int. J. of Humanities and Arts Computing **10**(1), 22–35 (2016)
21. Larson, R.: Bringing lives to light: Biography in context (2010), Final Project Report, University of Berkeley, http://metadata.berkeley.edu/Biography_Final_Report.pdf
22. Leskinen, P., Hyvönen, E.: Extracting genealogical networks of linked data from biographical texts. In: Proceedings of the 16th Extended Semantic Web Conference (ESWC 2019), Posters & demonstrations (June 2019)
23. Leskinen, P., Tuominen, J., Heino, E., Hyvönen, E.: An ontology and data infrastructure for publishing and using biographical linked data. In: Proceedings of the Workshop on Humanities in the Semantic Web (WHiSe II). CEUR Workshop Proceedings (October 2017)
24. Miyakita, G., Leskinen, P., Hyvönen, E.: Using linked data for prosopographical research of historical persons: Case U.S. Congress Legislators. In: 7th International Conference, EuroMed 2018, Proc., Part II. pp. 150–162. Springer-Verlag (2018)
25. Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., Bogaard, T.: Building event-centric knowledge graphs from news. Web Semantics: Science, Services and Agents on the World Wide Web **37**, 132–151 (2016)
26. Tuominen, J., Hyvönen, E., Leskinen, P.: Bio CRM: A data model for representing biographical data for prosopographical research. In: Biographical Data in a Digital World (BD2017) (2017), https://doi.org/10.5281/zenodo.1040712
27. Warren, C., Shore, D., Otis, J., Wang, L., Finegold, M., Shalizi, C.: Six degrees of Francis Bacon: A statistical method for reconstructing large historical social networks. Digital Humanities Quarterly **10**(3) (2016)