# Extracting Genealogical Networks of Linked Data from Biographical Texts

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

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

**Abstract.** This paper presents the idea and our work of extracting and reassembling a genealogical network automatically from a collection of biographies. The network can be used as a tool for network analysis of historical persons. The data has been published as Linked Data and as an interactive online service as part of the in-use system *BiographySampo – Finnish Biographies on the Semantic Web.*

## 1 Introduction

Extracting and inferring social or genealogical networks from historical documents can provide new information for biographical and prosopographical [1] research. However, genealogical data is often available only in textual form providing challenges for knowledge extraction: How to identify persons and their gender by different name forms? How to disambiguate namesakes in different times? How to extract the genealogical relations between the mentions? This paper presents a case study for extracting the explicit genealogical network implicit in the national collection of 13144 Finnish biographies[3]. The methodological idea is to combine regular expression identification, imprecise proper name matching, gender information, and data about expected lifespans for more accurate results. The system was evaluated with promising results, and a tool was constructed, based on Linked Data, for examining the underlying network of ∼81 000 extracted basic relations 'mother', 'father', 'wife', 'husband', 'son', and 'daughter'. On top of the Linked Data service, a new application was created for studying the networks interactively as a new part of the in-use BiographySampo[4] system [2].

**Related work** Extracting biographical networks are discussed in *Six Degrees of Francis Bacon* [3]. Articles [4,5] discuss extracting genealogical networks from multi-source vital records. For the large public there are many crowd-sourcing-based commercial genealogy websites, such as *ancestry.com*, *myheritage.com*, and *geni.com*. This paper extends our earlier papers about BiographySampo [2] and network analysis based on biographical link references into extraction of genealogical networks, and presents a new visual application view for studying such networks interactively.

[3] https://kansallisbiografia.fi/, accessed 20 March 2019
[4] http://biografiasampo.fi

## 2 Extracting Genealogical Networks from Texts

**Dataset** BiographySampo is a semantic portal based on a knowledge base that has been created using natural language processing methods, linked data, and semantic web technologies. It contains 13 144 biographies of notable Finns that can be browsed through a faceted search application and using tools for Digital Humanities research. [6] In addition to the genealogical network discussed in this paper, the data been a source for reference network extraction [7,8].

**Pattern-based Knowledge Extraction** Many biographies in the dataset include semi-formal textual descriptions of family relations of the protagonist. As an example, the description of baroness *Elisabeth Järnefelt*[5] is given below:

> Jelizaveta Konstantinovna Clodt von Jürgensburg from year 1857 known as Järnefelt, Elisabeth S 11.1.1839 Pietari, K 3.2.1929 Helsinki.
> *V* Baron, major general Konstantin Karlovitsh Clodt von Jürgensburg and Catharine Vign.
> *P* 1857 senator, governor, lieutenant general August Alexander Järnefelt *S* 1833, *K* 1896,
> *PV* bailiff Gustaf Adolf Järnefelt and Aurora Fredrika Molander.
> *Children:* Caspar (Kasper) Woldemar *S* 1859, *K* 1941, critic, translator, Russian language teacher, painter, *P* Emma Ahonen; Edvard Armas *S* 1869, *K* 1958, conductor, composer, professor, *P1* songstress Maikki Pakarinen, *P2* songstress Olivia (Liva) Edström; Aina (Aino) *S* 1871, *K* 1969, *P* composer Jean Sibelius;

The semi-formal expressions here have uniformity in structure that can be used effectively for pattern-based information extraction: First, the given and family names are mentioned and after that the years of birth $S$ and death $K$. The description provides information about the parents (marked with $V$), spouses ($P$), parents-in-law ($PV$), children, and children-in-law of the protagonist.

One major problem in knowledge extraction here is recognizing the same person, here *Elisabeth Järnefelt*, referenced with different names: *Jelizaveta Konstantinovna Clodt von Jürgensburg*, *Elisabeth Clodt von Jürgensburg* or most commonly with *Elisabeth Järnefelt*. On the other hand, same names are used in families over and over again. For example, there is a case of four people with name *Christian Trapp*, a grandfather, a father, a son[6], and a grandson. They cannot be distinguished without additional information about their known lifespans.

**Data Processing** In our knowledge extraction pipeline, the genealogical textual description of the protagonist is first divided into the parts describing his/her parents, spouses (wife/husband distinction is not known at this point), and children. The division is based on using regular expressions matching the punctuation and the tokens $V$, $P$, $PV$. The years of birth, death, or marriage are easily separated from the text sequence. To separate occupational descriptions from the proper names, we used the ARPA service[7] together with vocabularies of Finnish female, male, and family names[8].

The extracted names were used to reason the gender of the person, which was used to refine relations, e.g., to specify a *parent* as a *mother* or a *father*. For

[5] `http://biografiasampo.fi/henkilo/p3148`
[6] `http://biografiasampo.fi/henkilo/p10013`
[7] `http://seco.cs.aalto.fi/projects/dcert/`, accessed: 9 March 2019
[8] `https://www.avoindata.fi/data/en_GB/dataset/none`, accessed: 20 March 2019

the network, the spouses were linked with the children by the known years of marriage and child birth.

To gain detailed vital information for disambiguation, we reasoned lifetime estimates, e.g., the missing years of birth of the parents based on the known birth year of their child. The estimates were constructed by first collecting the years of births of a parent and a child from the known cases in data. The distributions of parent ages at child birth are depicted in Fig. 1. To reason the ages of spouses, a similar study was performed with the result that 99% of differences between the births of a husband and wife is in the range of -18–+35 years. The more relatives with known vital records a person has, the more precise the estimates are.

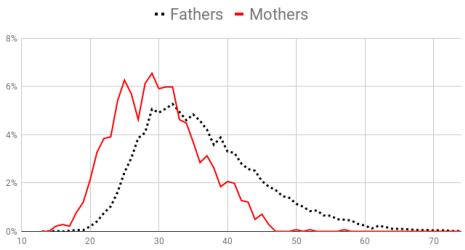


Fig. 1: Distribution of parent ages at a child birth

## 3 Evaluation

For evaluation we chose 50 random biographies, and manually compared the texts with the extracted results. The test set had mentions of 170 people. We compared the data fields of person names, years of birth and death, gender, occupation, and relation type. According to our evaluation 94.5% of the extracted people records were mentioned only in a single biography. The accuracy for these people was 97.3%, and 80.4% for people mentioned in multiple biographies. The system for inferring the gender could recognize 97.7% of the names leaving out very rare or foreign names. In our test set all inferred genders were correct.

For an example of the extracted network, the genealogical network of Elisabeth Järnefelt[9] is (partly) depicted in Fig. 2. She turns out to be a part of the largest connected subnetwork in our data. This subnetwork has 2694 family relation links and connects 1835 people mentioned in 250 biographies.

To further enrich the web portal, the network of immediate family members was used to reason other relatives of each protagonist[10]: the siblings, cousins, uncles, aunts, grandparents, grandchildren, and relatives-in-law.

[9] `http://biografiasampo.fi/henkilo/p3148/sukulaiset`
[10] `http://biografiasampo.fi/henkilo/p3148`

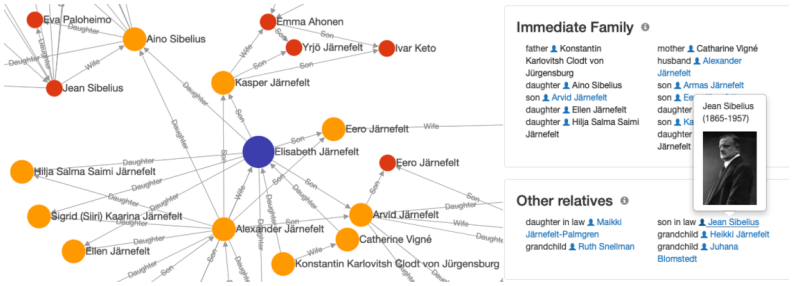


Fig. 2: Left: Genealogical network around Elisabeth Järnefelt as seen in a BiographySampo view. Right: relations shown on the web portal, including the composer Jean Sibelius, husband of Elisabeth's daughter Aino.

**Acknowledgements** Thanks to Business Finland for financial support and CSC – IT Center for Science, Finland, for computational resources.

## References

1. Verboven, K., Carlier, M., Dumolyn, J.: A short manual to the art of prosopography. In: Prosopography approaches and applications. A handbook. Unit for Prosopographical Research (Linacre College) (2007) 35–70
2. 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 ESWC 2019, Springer-Verlag (2019) Accepted.
3. Warren, C.N., Shore, D., Otis, J., Wang, L., Finegold, M., Shalizi, C.: Six degrees of Francis Bacon: A statistical method for reconstructing large historical social networks. DHQ: Digital Humanities Quarterly **10** (2016)
4. Efremova, J., Ranjbar-Sahraei, B., Rahmani, H., Oliehoek, F.A., Calders, T., Tuyls, K., Weiss, G.: Multi-source entity resolution for genealogical data. In: Population reconstruction. Springer-Verlag (2015) 129–154
5. Malmi, E., Rasa, M., Gionis, A.: AncestryAI: A tool for exploring computationally inferred family trees. In: Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee (2017) 257–261
6. Hyvönen, E., Leskinen, P., Tamper, M., Tuominen, J., Keravuori, K.: Semantic National Biography of Finland. In: Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018), CEUR Workshop Proceedings, Vol-2084 (2018) 372–385 `http://www.ceur-ws.org/Vol-2084/short12.pdf`.
7. Tamper, M., Leskinen, P., Apajalahti, K., Hyvönen, E.: Using biographical texts as linked data for prosopographical research and applications. In: Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. 7th International Conference, EuroMed 2018, Nicosia, Cyprus, Springer-Verlag (2018)
8. Tamper, M., Hyvönen, E., Leskinen, P.: Visualizing and analyzing networks of named entities in biographical dictionaries for digital humanities research. In: Proceeedings of CICLing 2019, Springer-Verlag (2019) Accepted.