Reconciling and Using Historical Person Registers as Linked Open Data in the AcademySampo Portal and Data Service

Petri Leskinen[1(✉)] and Eero Hyvönen[1,2]

[1] Semantic Computing Research Group (SeCo), Aalto University, Espoo, Finland
{petri.leskinen,eero.hyvonen}@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 a method for extracting and reassembling a genealogical network automatically from a biographical register of historical people. The method is applied to a dataset of short textual biographies about all 28 000 Finnish and Swedish academic people educated in 1640–1899 in Finland. The aim is to connect and disambiguate the relatives mentioned in the biographies in order to build a continuous, genealogical network, which can be used in Digital Humanities for data and network analysis of historical academic people and their lives. An artificial neural network approach is presented for solving a supervised learning task to disambiguate relatives mentioned in the register descriptions using basic biographical information enhanced with an ontology of vocations and additional occasionally sparse genealogical information. Evaluation results of the record linkage are promising and provide novel insights into the problem of historical people register reconciliation. The outcome of the work has been used in practise as part of the in-use AcademySampo portal and linked open data service, a new member in the Sampo series of cultural heritage applications for Digital Humanities.

Keywords: Data reconciling · Biographies · Linked data · Digital humanities

1 Introduction

A key idea of Linked Data is to enrich datasets by integrating complementary local information sources in an interoperable way into a global knowledge graph. This involves harmonization of local data models used, as well as aligning the concepts and entities used in populating the local data models. The latter problem has been addressed traditionally in the field of *record linkage (RL)* [7,13,36], where the goal is to find matching data records between heterogeneous databases. For example, how to match person records in different registers, which may contain data about same persons, but where the data is represented using different metadata schemas and notational conventions? Using RL, richer global descriptions of persons can be

A. Hotho et al. (Eds.): ISWC 2021, LNCS 12922, pp. 714–730, 2021.
https://doi.org/10.1007/978-3-030-88361-4_42

created based on fusing local datasets. In addition, RL facilitates data enrichment by linking together local datasets that use different vocabularies and identifiers for representing same resources, such as persons.

This paper concerns the problem of entity reconciliation and RL of people in historical person registers. As a case study, academic people and their relatives extracted automatically from the textual biographical descriptions of the Royal Academy of Turku and University of Helsinki are considered. The primary data contains some 28 000 short biographical descriptions of people in 1640–1899, covering virtually all university students in Finland during this time period. This data contains not only the 1) the explicit set of students recorded but also 2) the implicit set of persons mentioned in the short biography record texts of (1), such as relatives and prominent historical persons. The task is to construct a knowledge graph of all persons referred to in the data (1)–(2) in order to study the characteristics of the underlying academic network.

As a solution approach, a probabilistic RL solution for linking person records is presented and tested with promising evaluation results. In our method, RL is based on the attributes of an actor, such as the name, life years, and vocations relating to her/his life. The key novel idea here is to enrich these attributes with genealogical information, i.e., information about the names and lifespans of actors' relatives. Integrating local person registers into a single global *knowledge graph (KG)* facilitates biographical and prosopographical research based on enriched data. For this purpose, the aligned enriched person data has been used as a basis for a new in-use semantic portal and data service, *AcademySampo – Finnish Academic People 1640–1899 on the Semantic Web*[1]. The linked data model, data extraction, and data service of AcademySampo are described in [25], while in this paper the focus is on describing data reconciling and linking methods used, as well as on illustrating how the data service and semantic portal are actually used. More details on using the system (in Finnish) are available in [17].

This paper is structured as follows: We first present related works, the primary data of our study, and how it has been transformed into Linked Data. After this, the method of reconciling mentions of person in person registries is explained, and evaluation results in our case study are presented. In conclusion, contributions of the paper are discussed, and directions for further research are pointed out.

2 Related Work

The RL field is presented in [3,13,36]. Several nation-wide projects are underway on integrating person registries. For example, the Norwegian Historical Population Register (HPR) is pursuing to cover the country's whole population in 1800–1964, based on combining church records and census data [32]. The Links

[1] The portal and its linked open data service, including a SPARQL endpoint, was released on February 5, 2021. More information about AcademySampo can be found on the project homepage: https://seco.cs.aalto.fi/projects/yo-matrikkelit/.

project[2] in the Netherlands aims to reconstruct all nineteenth and early twentieth century families in the Netherlands based on civil certificates.

The problem of reconciling person records is evident in genealogical research. For example, in [26] Machine Learning has been applied to automatic construction of family trees from person records. Antolie et al. [2] present a case study of integrating Canadian World War I data from three sources: soldier records, casualty records, and census data. Here more traditional crafted RL processes were used, and using the data in research is demonstrated. Also Cunningham [8] concerns military person data. Here World War I military service records have been integrated with a census data, and the integrated data is used for data analysis. In Ivie et al. [19] the RL process is enhanced with the available genealogical data, e.g., information about spouses and children, to achieve a higher accuracy. Also Pixton et al. [27] utilize the genealogical information and apply a neural network for RL. 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 [4], BD2017 [11], and BD2019. In [23], analytic visualizations were created based on U.S. Legislator registry data, and the Six Degrees of Francis Bacon system[3] [22,34] utilizes data of the Oxford Dictionary of National Biography. Extracting Linked Data from texts has been studied in several works, such as [12]. In [10], language technology was applied for extracting entities and relations in RDF using Dutch biographies in the BiographyNet, as part of the larger NewsReader project [29].

Our own earlier works related to the topic include reconciling biographees and their relatives in the BiographySampo semantic portal [15,24]. Here genealogical statistics, e.g., average ages of becoming a parent or getting married were extracted from the source data, and person's life years are estimated according to that distribution. References to World War II soldiers were reconciled for data linking in the WarSampo portal and knowledge graph [14,21]. Unlike in these projects, in this paper a neural network model is trained to learn the classification rules from the existing ground truth linkage.

3 Knowledge Graph of Historical Academic Persons

This section presents the data used in our study: the Finnish university student registries "Ylioppilasmatrikkeli" containing short biographical descriptions.

3.1 Primary Data Sources

The student registry datasets in our focus are based on original handwritten university enrollment documents. In an earlier project, the documents have been transliterated manually into textual form and extended with information from

[2] Cf. the project homepage https://iisg.amsterdam/en/hsn/projects/links and research papers at https://iisg.amsterdam/en/hsn/projects/links/links-publications.
[3] http://www.sixdegreesoffrancisbacon.com.

other sources about later life events of the biographees. It has been estimated that ten man years of manual work of archivists was needed to accomplish this.

Our work concerns two main parts of the student registry: the database covering the years 1640–1852[4] available in Finnish and Swedish, and the registry of 1853–1899[5] for the next years. The records contain short biographical descriptions of 28 000 students of the University of Helsinki[6], originally the Royal Academy of Turku[7] in Finland. These student registries cover 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. The data is widely used by genealogists and historians. There are lots of mentions of relatives as well as of prominent related persons in the biographical descriptions. Generally, the data is divided into four parts: the students (*D1640*) in 1640–1852 register and their relatives (*R1640*) and likewise the students (*D1853*) and their relatives (*R1853*) in the later register.

A key challenge in transforming this kind of data into Linked Data for data-analysis is how to reconcile mentions of people in the records and their biographical texts. For example, the data contains records of ten students with the same name *Johan Wegelius*. In addition, eight of them have a vocation related to clergy—more than half of the students who studied before the year 1780 worked as priests after their graduation.[8] In the textual descriptions of the students, there are 72 mentions of spouses or mothers with the name *Maria Johansdotter*. Furthermore, there are variations in how the names are written because the data has been collected from multiple sources by different archivists, when it was extended by additional information about the later lives of the students. For example, the name *Sofia Dorotea Cedercreutz* can also be written as *Sophia Dorothea Cedercreutz*.

3.2 Extracting Information from Text

A comprehensive description about the data conversion as well as about the used data model is presented in an earlier article [25]. For example, an extract of the registry entry for *Anders Israel Cajander*[9] is depicted in Fig. 1. The description starts with the date or year of enrollment, in this case *11.2.1830*. After that there is the full name and a unique database identifier followed by the place and time of birth (Leppävirralla 24.2.1811). 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, one related to studies and academic career, and other describing the later career

[4] https://ylioppilasmatrikkeli.helsinki.fi.
[5] https://ylioppilasmatrikkeli.helsinki.fi/1853-1899.
[6] https://en.wikipedia.org/wiki/University_of_Helsinki.
[7] https://en.wikipedia.org/wiki/Royal_Academy_of_Turku.
[8] This statistical result was obtained after we used the reconciled data in AcademySampo for data analysis.
[9] https://ylioppilasmatrikkeli.helsinki.fi/henkilo.php?id=14689.

of the biographee. At the end of the first paragraph, a person's death is marked with the symbol † and burial with ‡; the person in the example died in Wyborg on December 18th, 1901 († Viipurissa 18.12.1901).

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).

Fig. 1. Partial extract from a register entry text for *Anders Israel Cajander*

After the life time description, there are possible fields for relatives. In the example case, the spouse is mentioned first as 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, i.e., two brothers (Veli: Gustaf Adolf Cajander and Veli: Zakarias Cajander) and a brother-in-law (Lanko: Berndt Vilhelm Kristoffer Schildt). The author of the *D1640* dataset, Yrjö Kotivuori, has manually added links from the description texts to the mentioned people also found in the register, like the three relatives in the example case. These links also contain linkage to the relatives in the *D1853* dataset.

3.3 Available Information

The previous person example was from the *D1640* data. However, the provided data in *D1853* differs in some aspects. For instance, *D1853* only mentions a person's parents and spouses, never children or any other relatives, and the people are not interlinked. Abbreviations are used generally for, e.g., vocations, which was taken into consideration in the data conversion by using specific lists of abbreviations.

Generally, the record linkage consists of the following partial tasks: 1) linkage from *R1640* to *D1640*, 2) linkage from *R1853* to *D1853*, 3) linkage from *R1853* to *D1640*, and 4) disambiguation of *R1640* and 5) *R1853* data. Table 1 shows an analysis of the known positive sample pairs in the both datasets. Here column *source* refers to the relative, and *target* to the corresponding student entry. The rows show how many of the example pairs of particular data field are available, altogether the data contains 4285 training pairs. One can notice that for the six uppermost properties, e.g., preferred label, gender, death, vocation, child, and spouse are available for both the source and target records. On the other hand, the data fields indicating the place of death, year of birth, names of mother or father, as well as the alternative labels are usually not available. The column *common* indicates the number of cases where both the source and the target entries have the particular data field and *same* the number of entries where the source and the target values are equal.

This table clearly indicates which properties should be considered crucial in decision making. Notice that some attributes that are usually significant for a general case of RL, such as places of birth and death, are not chosen in this particular case study.

Table 1. Available data fields in the training data

	Data 1640–1852				Data 1853–1899			
	Source	Target	Common	Same	Source	Target	Common	Same
Preferable label	4285	4285	4285	3979	698	698	698	517
Gender	4283	4284	4283	4283	698	698	698	688
Year of death	4229	4208	4192	4141	135	352	134	130
Vocation	4281	4270	4270	940	600	567	543	365
Child	4285	4284	4284	3211	430	341	340	2
Spouse	4285	4273	4273	–	698	687	687	2
Place of death	2	3494	2	2	–	348	–	–
Year of birth	–	2906	–	–	–	351	–	–
Mother	–	3475	–	–	–	349	–	–
Father	–	3478	–	–	–	348	–	–
Alternative label	–	1761	–	–	30	165	29	22

4 Method: Linking Person Records

This section describes the chosen formats for comparing two person registry entries. Generally, the input format for data comparison consists of numeric difference or similarity values between the data points of the two records, not the data of the records as it is. We first introduce the chosen input formats for data in different domains, e.g., for names and for vocations of the actors and the relatives. Finally, the architecture of the network model as well as the training setting are introduced.

4.1 Person Names

Person names in the datasets consist of a preferred and possibly alternative labels. Each label includes a family name and a sequence of given names. For the classifier input we considered four different variations of a label with a maximum of three first given names, only 0.4% of people entries have more than three given names. The classifier input is in a matrix format where the entry elements are statistical values calculated from the dataset. Each family and given name gets a *rarity* value so that first the frequency of the appearances for each name is counted and the ranks are mapped into the numeric range $[0.0, 1.0]$: the most common names get a near-zero and the rarest values closer to 1.0 in order to distinguish the rare names.

Figure 2 depicts an example of a name comparison matrix, in this case the family names of two person entries. The rows and columns mutually correspond to the data of two names that are compared. The uppermost row $(0.000, 0.808, 0.983, ...)$ consists of the rarity values for the first, and likewise the leftmost column $(0.000, 0.987, 0.991, 0.100, ...)$ for the second entry. The other values inside the matrix are Jaro-Winkler similarity values [35] between the name strings so that e.g., perfectly matching names get the value 1.0.

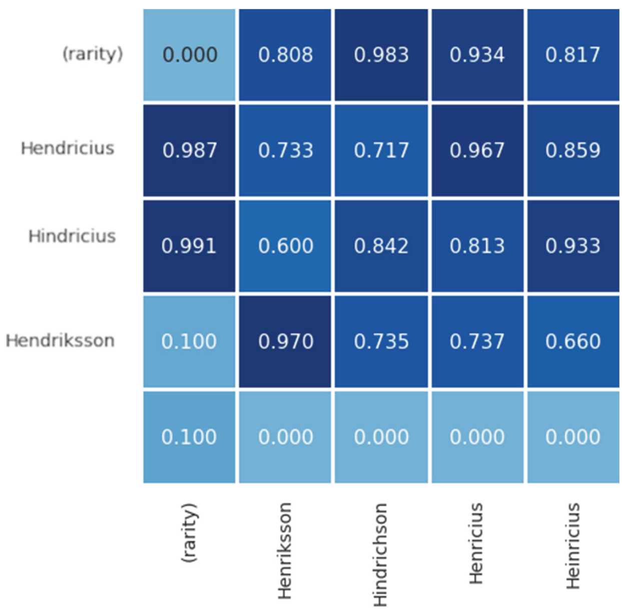

Fig. 2. Example of a matrix for comparing family names

4.2 Vocations

The vocations are the titles extracted from the source data. These titles often consists of a place name and a related profession, e.g., *Bishop of Turku* or *Bishop of Porvoo*. To enrich the data the vocations are linked to the hierarchical AMMO [20] ontology of historical occupations. Statistical values are used here like with the name entries. A *rarity* value is calculated for each title following the same principle as with the titles. In addition to that, a value of co-occurrences between two titles is calculated.

Figure 3 depicts an example of a vocation comparison matrix. The value in the leftmost upper corner (0.455) is the Jaccard index [31] between the two sets of vocations. Similarly to the name matrix, the rows and columns correspond to the vocation in two dataset entries with the rarity values on the uppermost row $(0.909, 0.804, 0.249...)$ and leftmost columns. The rarity values are in a descending order so that the rarest vocations appear first on the lists. The other values filling the rest of the matrix are the co-occurrence values. In the data matrix, the co-occurrence value for a pair *(Law Reader, Mayor)* is 0.985, while the pair *(Court Attorney, Mayor)* has a value 0.250 indicating that this pair co-occurs in the data more frequently. The zero-valued elements on the right indicate that one of the title sets has less than the reserved seven data fields.

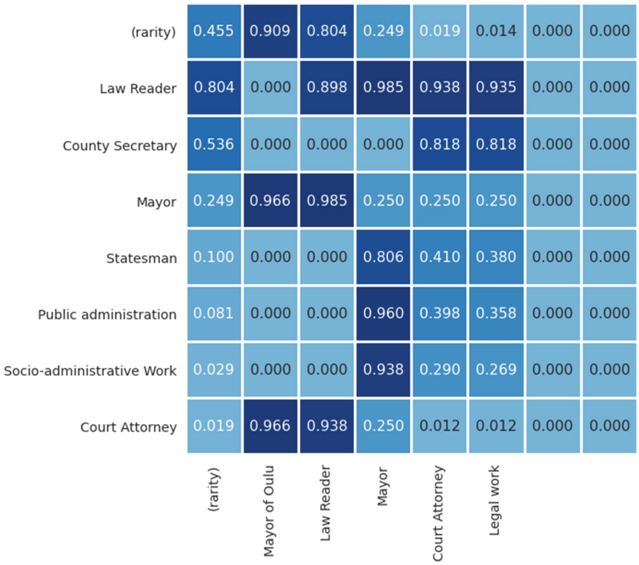

Fig. 3. Matrix for comparing the vocations

4.3 Years of Birth and Death, Gender

The difference in actor's and relatives' birth and death years and their genders were also input to the network. The years use a precision of one year due to the format used in source data: the birth and death of the actor is usually known with a precision of a day, while in the case of relatives only the precision of a year is used. The actual difference is years is mapped into a near-zero range by using the arctan function. Gender was indexed using value -1.0 for female, 1.0 for male, and 0.0 for the rare cases where the gender was not known.

4.4 Relative Information

The information of the relatives consists of details about the children and spouses of an actor, and basic information about her/his parents. The relative information uses the same matrix format as for the names, lifetime information, and vocations of the actor. It has reserved space for three children and three spouses, according to analyzing the data. In the data more than 99% have three or less spouses, and 95% three or less children.

4.5 Network Model

The used network model is depicted in Fig. 4. It is a multi-input network based on the Keras functional API [6]. The network has eight inputs out of which six for the given and family names of the two actors, their spouses, and their children, one for the age comparison of actors and their relatives, and one for the actors' titles. The network acts as a probabilistic classifier and the output is $\bar{y} \approx [0.0, 1.0]$ for matching entries and $\bar{y} \approx [1.0, 0.0]$ for not matching pairs. For a binary decision these values are filtered by choosing the positive matches when the latter value exceeds a chosen threshold, e.g., $\lambda = 0.9$.

Some inputs are in a matrix format, which are first flattened[10], and after that run through a Dense[11] layer. Dropout layers with a ratio of 25% are used to prevent the overfit to the training data [30]. Different inputs of the same domain (e.g., names and years) are first concatenated[12] to one another. After a layer of Dense network the network concatenates into the final output.

Training Data. The training data for the neural network could be input by either as an single data entry or in several smaller batches of data. We chose to feed the data in batches utilizing the Keras Data Generator Sequence [1] as described by A. Amidi and S. Amidi[13] due to the amount of data preprocessing from RDF format to numeric input.

Positive samples are created by reading the manually marked matches from the data. This linkage is many-to-one, so all the samples pointing to the same target can be chosen as training pairs pointing to each other. Finally, positive sample data is augmented with pairs where both the target and the source refer to the same resource.

The easiest way to gather the negative samples is to pick random pairs from the data. However, we chose to sample pairs that are likely to have some similar data values to improve the decision making. The dataset contains relations indicating, e.g., that two persons are siblings, cousins, or namesakes. Close relatives often have same similar characteristics, like family name or nearby years of birth.

[10] https://keras.io/api/layers/reshaping_layers/flatten/.
[11] https://keras.io/api/layers/core_layers/dense/.
[12] https://keras.io/api/layers/merging_layers/concatenate/.
[13] https://stanford.edu/~shervine/blog/keras-how-to-generate-data-on-the-fly.

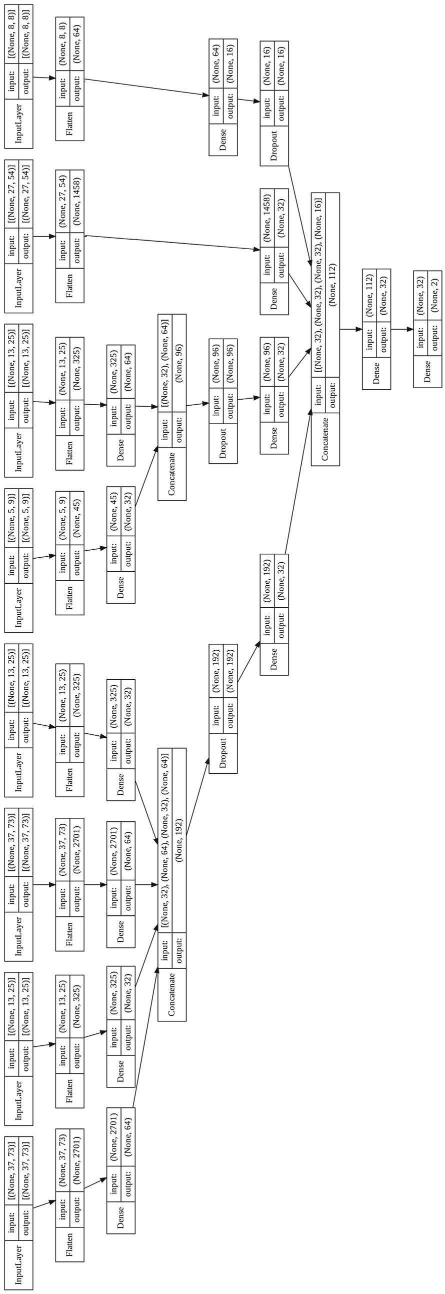

Fig. 4. Classifier model structure

Model Training. For the training the data was split into separate sets for training, testing, and validation of sizes 70%, 15%, and 15%, respectively. The classes in the training data are imbalanced, e.g., the number of negative samples ($N_n \approx 200000$) is significantly larger than the positive samples ($N_p \approx 13000$). Therefore the positive samples where defined to have a larger weight than the negative ones [5,33]. The training was performed in Google Colab, and the training with 100 epochs using a GPU took 4242.2 s. Validation accuracy of more than 99.6% was achieved during the training.

4.6 Evaluation

The results were analyzed closely by the Receiver Operating Characteristic (ROC) curve (Fig. 5) and by taking look at the details of False Positive and False Negative classifications. To deal with the data imbalance, a validation set with equal amount of positive and negative sample was used. The classifier input was divided by four different types: basic biographical information (B), genealogical information (G), name frequencies (N), and vocation frequencies (V). To analyze how much each data entry contributes to the prediction, evaluation was performed for four times using the entire data (B+G+N+V), biographical and genealogical data (B+G), biographical data with name and vocation frequencies (B+N+V), and the plain biographical data (B). The threshold value λ for optimal performance was chosen from the ROC curve coordinates by the point closest to the upper left corner [9]. For the entire data (B+G+N+V) the threshold value was $\lambda = 90.01\%$ and the resulting number of True Positives (TP) is 2035, True Negatives (TN) 2089, False Positives (FP) 0, and False Negatives (FN) 54 with measures precision of 100.00%, recall of 97.42%, F_1-score of 98.69%, and accuracy of 98.71%.

In the ROC visualization, the curve with basic and genealogical (B+G) almost emerges with the curve for the entire data (B+G+N+V). Also Table 2 shows how close these results are to one another. Furthermore, the validation results without the genealogical information (B, B+N+V) show lower accuracy.

Table 2. Validation results using different data subsets

Data Subset	TP	FP	FN	TN	Precision	Recall	F_1-score	Accuracy	AUC	λ
B+G+N+V	2035	0	54	2089	100.00%	97.42%	98.69%	98.71%	99.98%	90.01%
B+G	2007	1	82	2088	99.95%	96.07%	97.97%	98.01%	99.97%	84.06%
B+N+V	2011	150	78	1939	93.06%	96.27%	94.64%	94.54%	98.47%	16.86%
B	587	12	1502	2077	98.00%	28.10%	43.68%	63.76%	97.48%	92.15%

Full Disambiguation. Record linkage with the real dataset was a many-to-one task, e.g. many records in the source set can be merged into one in the target data. When applying the model to the real dataset first blocking strategies [7]

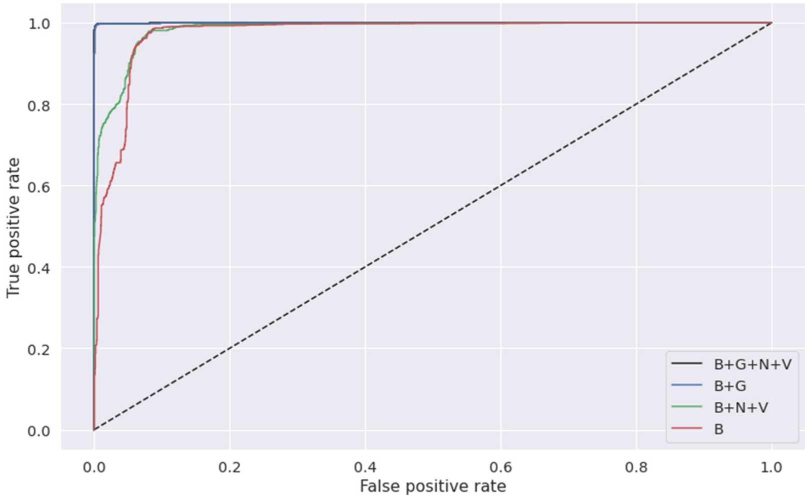

Fig. 5. ROC curve

where applied to reduce the number of comparisons. For instance, candidate pairs of different gender or mismatching life years when known, could be omitted from candidate pairs. Likewise, candidates mentioned in a same register entry text e.g. siblings or different spouses could be omitted—same person is never mentioned twice in one text entry. Some preliminary disambiguation was performed already during the data conversion, e.g., aligning spouses of a person, if the names had a high string similarity. The iterative process was run for several times because merging two person records furthermore can lead to finding more matches also among the relatives. To achieve a high precision and to minimize the number of false positive classification a high threshold values ($\lambda \geq 0.9$) were used.

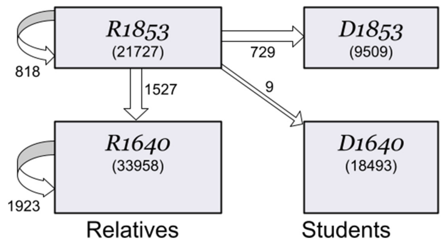

Fig. 6. Number of matches between the datasets

Figure 6 depicts the number of records in each part of the dataset and the numbers of matches detected within them. The number of records before the RL are in parenthesis. For example, 729 of the records in *R1853* were merged into *D1853*, 1527 into *R1640*, and 9 records into *D1640*. The latter number is relatively small because this matching was a part of the existing manual linkage by the dataset author, so these results are links missing from manual linkage or errors in our data conversion process. Inside the *R1853* dataset, 818 and in *R1640* 1923 entries were matched, respectively. Notice that we did not link the records from *R1640* to *D1640* because the existing manual linkage made by the dataset author.

5 Using AcademySampo

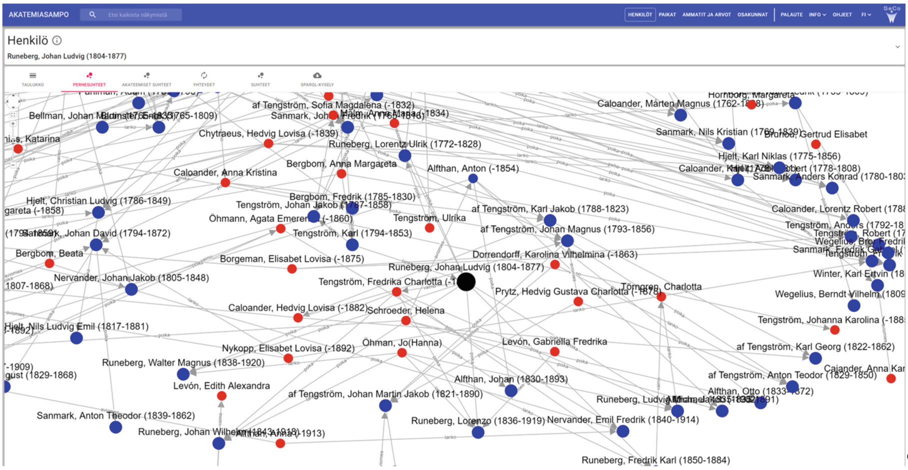

Fig. 7. Family relations of J. L. Runeberg (1804–1877) visualized in AcademySampo (Color figure online)

The people KG extracted from the primary data turned out be richly interlinked and forms the backbone of the AcademySampo portal and LOD service. Academic circles in history were smaller and people tended to marry within their own social class. For example, Fig. 7 depicts the extracted family relations of J. L. Runeberg (1804–1877) (black large spot in the centre), the Finnish national poet, as visualized in one of the data-analytic views of the AcademySampo portal. Men in the figure are represented as blue and women as red spots. Most women in the data do not have a data entry of their own in the databases but are only mentioned in the biographies of the men because women were allowed to sign in universities only in the late 19th century. There are only 521 female academics out of 28 000 in the data.

The relations shown include both mentioned and inferred relations, such as brother in law, based on reasoning. Here is an example[14] of a SPARQL query that finds children of the same parent and concludes whether they are brothers

[14] https://api.triplydb.com/s/IE4w29n0T.

or sisters based on the gender. Using AcademySampo portal and the SPARQL endpoint for historical research is discussed in more detail in [17].

Deployment. The AcademySampo KG was published on the Linked Data Finland platform[15] [16] powered by Fuseki SPARQL server[16] and Varnish Cache web application accelerator[17] for routing URIs, content negotiation, and caching. The portal user interface was implemented by the Sampo-UI framework [18]. AcademySampo system is based on Docker microservice architecture containers[18]. By using containers, the services can be migrated to another computing environment easily, and third parties can re-use and run the services on their own. The architecture also allows for horizontal scaling for high availability, by starting new container replicas on demand. The portal has had 4600 distinct users during its first four months according to Google Analytics.

6 Discussion

The work described in this article shows that using genealogical information in RL is useful and can improve significantly the accuracy in person name reconciliation. This argument was tested and evaluated in detail in a case study using the AcademySampo datasets with promising results. We anticipate that similar results can be obtained in related use cases using other datasets. In the AcademySampo project, the genealogical information has been used also when linking the records with Wikidata for semantic data enrichment.

When analysing the resulting matched pairs some weak cases needing separate handling where found. Historically, patronymic family names, e.g., *Johansdotter (Daughter of Johan)* have been common for women. However, the chosen Jaro-Winkler similarity may not be optimal to always disambiguate between cases like *Jöransdotter* and *Johansdotter*. Likewise, the classifier made some false results with the vocation of a farmer. Farmer was a common vocation in the 17th–19th century Finland, but yet rare in data records of academic people, for which reason we had put some excess weight on it in the classifying system.

This paper presented a method for reconciling person names mentioned in biographical texts of other people. The method was applied to creating a semantic KG of people that is used for studying and analyzing academic networks of people. For this purpose, the AcademySampo portal has been created, but also the underlying Linked Open Data service can be used for custom-made data-analyses using, e.g., YASGUI[19] [28] and SPARQL or Python scripting in Google Colab[20] or Jupyter[21] notebooks, and for developing new applications [17].

---

15 https://www.ldf.fi/dataset/yoma.
16 https://jena.apache.org/documentation/fuseki2/.
17 https://varnish-cache.org.
18 https://www.docker.com.
19 https://yasgui.triply.cc.
20 https://colab.research.google.com/notebooks/intro.ipynb.
21 https://jupyter.org.

Acknowledgements. Thanks to Yrjö Kotivuori and Veli-Matti Autio for their seminal work in creating the original databases used in our work, and for making the data openly available. Discussions with Heikki Rantala, Esko Ikkala, Mikko Koho, and Jouni Tuominen are acknowledged. This work is part of the EU project InTaVia: In/Tangible European Heritage (https://intavia.eu/), and is related to the EU COST action Nexus Linguarum (https://nexuslinguarum.eu/the-action) on linguistic data science. CSC – IT Center for Science provided computational resources for the work.

References

1. Keras Documentation, Sequence. https://www.tensorflow.org/api_docs/python/tf/keras/utils/Sequence. Accessed 10 Dec 2020
2. Antonie, L., Gadgil, H., Grewal, G., Inwood, K.: Historical data integration, a study of WWI Canadian soldiers. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 186–193. IEEE (2016)
3. Barlaug, N., Gulla, J.A.: Neural networks for entity matching. arXiv preprint arXiv:2010.11075 (2020)
4. 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/
5. Brownlee, J.: Machine Learning Mastery: How to Develop a Cost-Sensitive Neural Network for Imbalanced Classification. https://machinelearningmastery.com/cost-sensitive-neural-network-for-imbalanced-classification/. Accessed 10 Dec 2020
6. Chollet, F.: Keras, The Functional API. https://keras.io/guides/functional_api/. Accessed 10 Dec 2020
7. Christen, P.: Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31164-2
8. Cunningham, A.: After "it's over over there": using record linkage to enable the reconstruction of World War I veterans' demography from soldiers' experiences to civilian populations. Historical Methods: J. Quant. Interdisc. Hist. **51**, 1–27 (2018)
9. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
10. Fokkens, A., et al.: BiographyNet: extracting relations between people and events. In: Europa baut auf Biographien, pp. 193–224. New Academic Press, Wien (2017)
11. 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/
12. Gangemi, A., Presutti, V., Recupero, D.R., Nuzzolese, A.G., Draicchio, F., Mongiovì, M.: Semantic web machine reading with FRED. Semantic Web **8**, 873–893 (2017)
13. Gu, L., Baxter, R., Vickers, D., Rainsford, C.: Record linkage: current practice and future directions. CSIRO Mathematical and Information Sciences (2003). cMIS Technical Report No. 03/83
14. Heino, E., et al.: Named entity linking in a complex domain: case second world war history. In: Gracia, J., Bond, F., McCrae, J.P., Buitelaar, P., Chiarcos, C., Hellmann, S. (eds.) LDK 2017. LNCS (LNAI), vol. 10318, pp. 120–133. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59888-8_10

15. Hyvönen, E., et al.: BiographySampo – publishing and enriching biographies on the semantic web for digital humanities research. In: Hitzler, P., et al. (eds.) ESWC 2019. LNCS, vol. 11503, pp. 574–589. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-21348-0_37
16. 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.) ESWC 2014. LNCS, vol. 8798, pp. 226–230. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11955-7_24
17. Hyvönen, E., Leskinen, P., Rantala, H., Ikkala, E., Tuominen, J.: Akatemiasampo-portaali ja -datapalvelu henkilöiden ja henkilöryhmien historialliseen tutkimukseen (academysampo portal and data service for biographical and prosopographical research). Informaatiotutkimus (2021, in press). https://seco.cs.aalto.fi/publications/2021/hyvonen-et-al-akatemiasampo-2021.pdf
18. Ikkala, E., Hyvönen, E., Rantala, H., Koho, M.: Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. Semantic Web (2021, accepted). http://www.semantic-web-journal.net/
19. Ivie, S., Pixton, B., Giraud-Carrier, C.: Metric-based data mining model for genealogical record linkage. In: 2007 IEEE International Conference on Information Reuse and Integration, pp. 538–543. IEEE (2007)
20. Koho, M., Gasbarra, L., Tuominen, J., Rantala, H., Jokipii, I., Hyvönen, E.: AMMO ontology of Finnish historical occupations. In: Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH 2019), vol. 2375, pp. 91–96. CEUR Workshop Proceedings, June 2019. http://ceur-ws.org/Vol-2375/
21. Koho, M., Leskinen, P., Hyvönen, E.: Integrating historical person registers as linked open data in the WarSampo knowledge graph. In: Blomqvist, E., et al. (eds.) SEMANTICS 2020. LNCS, vol. 12378, pp. 118–126. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59833-4_8
22. Langmead, A., Otis, J., Warren, C., Weingart, S., Zilinski, L.: Towards interoperable network ontologies for the digital humanities. Int. J. Humanit. Arts Comput. **10**(1), 22–35 (2016)
23. Larson, R.: Bringing lives to light: biography in context. Final Project Report, University of Berkeley (2010). http://metadata.berkeley.edu/Biography_Final_Report.pdf
24. Leskinen, P., Hyvönen, E.: Extracting genealogical networks of linked data from biographical texts. In: Hitzler, P., et al. (eds.) ESWC 2019. LNCS, vol. 11762, pp. 121–125. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32327-1_24
25. Leskinen, P., Hyvönen, E.: 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, vol. 2612, pp. 284–292. CEUR Workshop Proceedings, October 2020. http://ceur-ws.org/Vol-2612/short14.pdf
26. Malmi, E., Gionis, A., Solin, A.: Computationally inferred genealogical networks uncover long-term trends in assortative mating. arXiv (2018). arXiv:1802.06055 [cs.SI]
27. Pixton, B., Giraud-Carrier, C.: Using structured neural networks for record linkage. In: Proceedings of the Sixth Annual Workshop on Technology for Family History and Genealogical Research (2006)
28. Rietveld, L., Hoekstra, R.: The YASGUI family of SPARQL clients. Semantic Web **8**(3), 373–383 (2017). https://doi.org/10.3233/SW-150197

29. Rospocher, M., et al.: Building event-centric knowledge graphs from news. Web Semantics **37**, 132–151 (2016)
30. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
31. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to data mining, 1st edn (2005)
32. Thorvaldsen, G., Andersen, T., Sommerseth, H.L.: Record linkage in the historical population register for Norway. In: Bloothooft, G., Christen, P., Mandemakers, K., Schraagen, M. (eds.) Population Reconstruction, pp. 155–171. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19884-2_8
33. Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., Kennedy, P.J.: Training deep neural networks on imbalanced data sets. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 4368–4374. IEEE (2016)
34. 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. Digit. Humanit. Q. **10**(3) (2016)
35. Winkler, W.E.: String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage (1990)
36. Winkler, W.E.: Overview of record linkage and current research directions. Technical report, U.S. Census Bureau (2006)