EasyChair Preprint

№ 888

# Visualizing and Analyzing Networks of Named Entities in Biographical Dictionaries for Digital Humanities Research

Minna Tamper, Eero Hyvönen and Petri Leskinen



April 8, 2019

# Visualizing and Analyzing Networks of Named Entities in Biographical Dictionaries for Digital Humanities Research

Minna Tamper[1 [0000−0002−3301−1705]], 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
[2] 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 shows how named entity extraction and network analysis can be used to examine biographies individually and in groups to aid historians in biographical and prosopographical research. For this purpose a reference network of 13 100 biographies in the collections of the Biographical Centre of the Finnish Literature Society was created, based on links between the biographies as well as automatically extracted named entities found in the texts. The data was published in a SPARQL endpoint as a Linked Data knowledge graph on top of which network analytic tools were created and analysis were done showing the usefulness of the approach in Digital Humanities. The reference graph has been utilized for network analysis to examine egocentric networks of individual persons as well as networks among groups of people in prosopography. The data and tools presented are in use since autumn 2018 in the semantic portal BiographySampo that has had tens of thousands of users.

## 1 Introduction

BiographySampo[3] is a semantic portal that is based on a knowledge graph that has been created using natural language processing methods, linked data, and semantic web technologies [15,35]. The graph contains currently ca. 13 100 biographical textual descriptions of notable Finns that can be browsed through using faceted search and a variety of data-analytic tools. 9200 of the entries contain a short, free text biography of the person, created by 977 professional authors. The portal has been built to help historians and scholars in biographical [33] and prosopographical research [37,8][4]. A major novelty of BiographySampo is to provide the user with data-analytic and visualization tools for solving research problems in Digital Humanities (DH), based on Linked Data [9,12].

[3] Online at www.biografiasampo.fi; see project homepage https://seco.cs.aalto.fi/projects/biografiasampo/en/ for further info and publications.

[4] Prosopography is a method that is used to study groups of people through their biographical data. The goal of prosopography is to find connections, trends, and patterns from these groups.

In the biography texts, the authors mention other people they consider significant from an occupational or other relevant perspective. In our case study, the editors of the dictionary of biography at the publisher Finnish Literature Society (SKS) have changed these mentions into internal links to corresponding articles in the dictionary if there is one.[5] A link is added typically only once when a person is mentioned for the first time. These links serve in the original biography collection as a way to browse and move between the biographies.

However, many links are missing from the text. For example, there are mentions of relatives and external people who do not have a biography in the dictionary to be linked to, e.g., William Shakespeare and Richard Wagner. In addition, if a biography $A$ mentions person $B$, but the biography of $B$ has been added in the collection after editing $A$, it has not been possible to add the link. The explicit links between people in the biographical texts therefore create a scarcely interlinked reference network of the biographical texts.

This paper argues that making the reference network underlying a biographical dictionary explicit can be useful in biographical and prosopographical research. The idea of using the network analysis of historical people for Digital Humanities research has been suggested before in, e.g., [38,3]. A contribution of our paper is to apply the idea to biography collections, where connections are based on entity mentions. To support the argument, we present a case study using BiographySampo where the reference network underlying its textual biographies was extracted and enriched into a knowledge graph and published as a linked data service, on top of which a set of tools were created for Digital Humanities research. This idea is currently being applied also to a genealogical network extracted from the same texts [20].

In the following, the underlying knowledge graph with its person and place ontologies, and the process of extracting and enriching the reference network is first presented (Section 2). After this, application views to study the networks underlying the biographical texts are presented (Section 3). Firstly, a network analysis tool is presented for visualizing and studying the *egocentric network* of a protagonist in biographical research. Secondly, this idea is generalized for prosopography where groups of people sharing characteristics (such as occupation, gender, or area of living) are studied. Here the user can first separate the target group using faceted search and then visualize the group's *sociocentric network*. Thirdly, when visualizing the networks, it turned out that they often include serendipitous [1] (surprising) connections between people, raising the question: why are these two people interconnected? A tool is clearly needed for explaining the connections, not only showing them. For this purpose, an application view showing the textual contexts in which the connections arise was created. Lastly, the toolset presented also includes an application called contextual reader [24], where the user is able to get information about the extracted linked entities by hovering the mouse on top of the mentions. After presenting the application

[5] Actually, the biographies in our case study come from several separate databases, including the general National Biography of Finland as a core, supplemented with four other thematic dictionaries [16].

views, the applications and named entity extraction is evaluated (Section 4). In conclusion (Section 5), the contributions of the paper are summarized, related works discussed, and the directions of further research suggested.

## 2 Extracting Named Entities from Biographical Texts

In order to build and integrate network analysis tools, reference analysis tools, and the contextual reader application to BiographySampo, the existing links and named entities need to be extracted from the texts, and the underlying BiographySampo Knowledge Graph (BSKG) be enriched accordingly. In this section the knowledge base, extraction process, and the data transformations that enable the end user applications are discussed.

**Knowledge Graph** BSKG includes the biography collections[6] of SKS written by 977 scholars from different fields. The biographies describe the lives and achievements of historical and contemporary figures, containing vast amounts of references to notable Finnish and foreign figures and to historical events, works (such as paintings, books, music, and acting), places, organizations, and dates. The graph includes 13 144 people with a biographical description, 51 200 related people mentioned in the biographies, and the 977 authors of the biographies. There are furthermore 225 000 lifetime events of the protagonists including their births, deaths, and other biographical events. The biographical texts also contain manually added 31 500 HTML links links between the biographs that were included in the knowledge base [35]. There is also a separate graph of 4970 places, extracted from the Finnish Gazetteer of Historical Places and Maps (Hipla) and data service[7] [17,14]. Foreign placenames were linked using the Google Maps APIs[8]. The lifetime events have lots of mentions of other kinds, such as governmental or educational buildings, public places etc. An additional dataset of approximately 2000 resources was extracted for them from Wikidata. The data was also augmented with a list of countries in the world and their capitals. [21]

**Extraction and Linking Process** The biographical texts [35] were transformed into an RDF dataset and enriched with linguistic information, totaling in 120 million triples. The data can be queried from a SPARQL endpoint. This data contains manually annotated links that have been extracted from the HTML as well as links based on entity linking.

Named entity linking tools [26,28,25] typically use a process that can be broken into three tasks [7,4]: 1) named entity recognition (NER), 2) named entity disambiguation (NED), and 3) named entity linking (NEL). NER identifies the entities from text, NED disambiguates them, and lastly NEL links the mentions to their meanings in ontologies or knowledge bases. Our new linking tool, Nelli

[6] `https://kansallisbiografia.fi/english/national-biography`
[7] `http://hipla.fi`
[8] `http://developers.google.com/maps/`

extracts and links entities from texts in a similar manner. However, in addition it combines multiple, in our case three different tools for NER and NEL. The purpose of this approach is to improve disambiguation by utilizing a voting scheme [32,6] where each tool has a vote on the interpretation it makes for the same piece of text. The best candidate is the one with most votes. For example, to identify a place from the string *Turku Cathedral* the tools return three answers of which one is *Turku* and two are *Turku Cathedral*, the winning interpretation.

Nelli uses the tools FiNER, ARPA, and LINFER. FiNER[9] is a rule-based NER tool for Finnish, ARPA [23] is a NER and NEL tool [18,10] that queries matches from controlled vocabularies. To supplement FiNER and ARPA, a third tool LINFER was implemented utilizing the linguistic RDF data to identify named entities. The parsed linguistic data not only contains part of speech information but also Dependency Grammar relations. With this information a set of rules was created to infer which proper nouns (or nouns) would be most likely placenames or person names. With this tool entities such as *Åbo Akademin kirjasto* (engl. *Åbo Akademi University Library*) can be identified by analyzing inflected forms and dependencies. These rules were encapsulated in LINFER to utilize the linguistic features of words and their relations.

In addition to each tool having a vote, votes can be earned for entity length, linkage, and by named entity type. Sometimes it may be difficult to identify correctly longer named entities, such as place or organization names, and therefore a vote is given to the longest matching candidates. Also candidates that have found a match in an ontology are favored with a vote. Nelli also has a priority order for named entity types where votes can be added to favor some entity types over others. For example, the address *Konemiehentie 2, Espoo* contains a name of a city. In order to have the address as the top voted candidate, it will help to give to the address type a higher score than to the more general location.

Once the Nelli has all the interpretations and metrics about the candidates, it calculates the votes and writes the results in Turtle format. For this extension of the original data, we used NLP Interchange Format (NIF)[10] [11], Dublin Core Metadata[11], and a custom namespace[12] to supply classes and properties that describe named entity metadata. For recording the results, the application writes *nbf:NamedEntity* class instances that have the basic information about the entity. It has properties to describe the extracted string (*nif:isString*), base form of the string (*nif:lemma*), its named entity type (*nbf:namedEntityType*), where it is linked (*skos:relatedMatch*), the location of the string in text (*nif:beginIndex*, *nif:endIndex*), and the method that was used to extract the named entity (*nbf:usedNeMethod*). In the source dataset, the texts have been split to documents, paragraphs, sentences, and words. The word-entities are also added a *dct:isPartOf* property referring to the named entity instances they are a part of and similarly the sentences have a *nbf:hasNamedEntity* property. The value of

[9] `https://github.com/Traubert/FiNer-rules/blob/master/finer-readme.md`
[10] `http://persistence.uni-leipzig.org/nlp2rdf/specification/core.html`
[11] `http://dublincore.org/documents/dcmi-terms/`
[12] denoted with prefix `nbf`

the *nbf:usedNeMethod* property is an instance of the *nbf:NamedEntityType* class that is the description of the named entity type. The value of the *nbf:usedNeMethod* property is an instance of the class *nbf:NamedEntityMethod* that has provenance information about the tools used to extract the named entity. In addition to the *nbf:NamedEntity* class, there is also the *nbf:NamedEntityGroup* class that groups the entities in each sentence based on location and possible overlap. Each group has all members indicated with the property *nbf:member* and the top voted entity with *nbf:primary*.

**Reference Networks** Network analysis of people [38,3] is a set of methods that can be used to study social networks [30]. In our case, the networks were built from the HTML links and mentions of people in the biographies to create a reference network which is analogous to citation networks [34]. In a reference network, the nodes are people, and when a person $A$ is mentioned in the biography of $B$, a directed edge is added from $B$ to $A$. The edges are instances of the class nbf:Reference with properties for the source biography *nbf:source*, the mentioned person *nbf:target*, and the type of the reference as *nbf:ManualAnnotation* (for HTML links) or *nbf:AutomaticAnnotation* (for identified named person entities). The number of references to the target person in the source biography is declared as the value of the *nbf:weight* property which for manual HTML links equals to one.

The transformed network data can then be used in applications by quering the nodes, e.g., biographical details of people, and the edges, e.g., the links between people. Based on the data, the networks can be generated automatically for an individual or a group.

## 3 Applications

To test the potential of network analysis in biography and prosopography, a reference network was constructed based on HTML links in ca. 6100 of the 13 144 biographies, enhanced with additional edges from 400 biographies by Nelli This group was limited to politicians, writers, athletes, lutherans, artists, architects, and musicians because their biographies contain long textual descriptions.

The BSKG included entities for people and places extracted from texts. For place linking, we used the YSO Places ontology[13] of the Finnish Ontology Service Finto that contains contemporary place resources for municipalities, provinces, countries, and continents. The contemporary data was extended with the WarSampo place ontology [13] that includes historical Finnish places. A priority order was set for place and person entities so that more specific place names, for instance, have a higher score. Also, to avoid having people's first and last names mislabeled as places, person named entities were given a higher score.

With this setup, 33 120 entities were extracted and used as a basis for four application views presented next in this section.

[13] `https://finto.fi/yso-paikat/en/`

**1. Egocentric Networks** The egocentric networks are formed from people nodes, i.e., biographical details of people, and edges, i.e., the links between people. The networks are generated to the center of the screen and centered around one person, in this case the protagonist. On the left hand side of the user interface, there are network toggles that can be used to alter the layout of the network in the following ways: Firstly, the user can tamper the amount of nodes to be seen, i.e., limit the size of the network to be visualized. Secondly, the user can select to see the network built using the manual HTML links only, automatically extracted links, or both. In this way, the manual and automatically extracted links can be compared with each other. Thirdly, to emphasize the most significant nodes in the graph, the node size can be determined based on using four distance and centrality measures used in network analysis: distance to the protagonist, degree, in-degree, out-degree, or pagerank [27]. Fourthly, it is possible to color the person nodes based on the gender, occupational area, or distance to the protagonist. The network is generated based on the selected toggle options, and the automatic links option shows the edge weight based on how frequently the person is mentioned in the text.

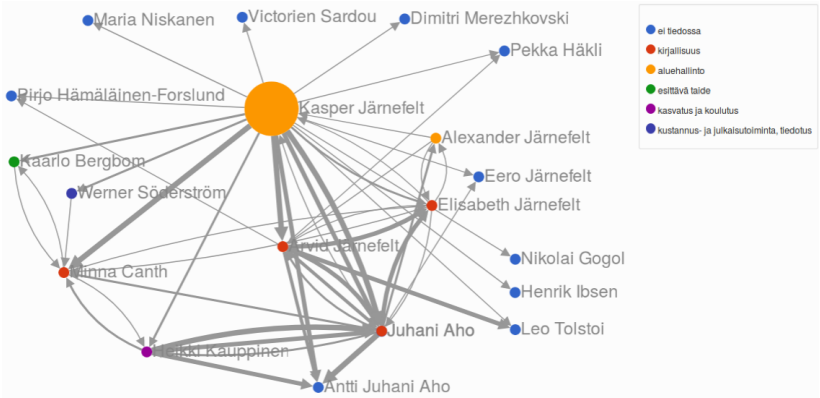


**Fig. 1.** Kasper Järnefelt's egocentric network where nodes are colored by occupation.

Fig. 1 depicts the egocentric network of the Finnish critic, translator, and cultural person Kasper Järnefelt (1859–1941). The network shows, e.g., lots of links to contemporary Finnish cultural persons with a biography in the system (based on the HTML links), as well as connections to external people, such as authors Nikolai Gogol, Henrik Ibsen, and Leo Tolstoi (based on Nelli), who do not have biography in BiographySampo. The linkage is based on the fact that Järnefelt has translated their works into Finnish. The width of the edges indicates the number of references between the biographies and is an indication of potential importance. The legend box in the right upper corner explains the color coding of the occupational areas used for the nodes. The toggles for making the selections for the visualization are not shown in the figure for brevity.

In BiographySampo, the egocentric networks are located under the *Network* tab in the personal home pages of the protagonists.

**2. Sociocentric Networks** The sociocentric networks are located in their own view in BiographySampo. They can be accessed from the navigation bar under the title *Verkostot* (engl. Networks). In this application view, the user first filters the target group she is interested in studying by using faceted search. For example, people of similar occupation or place of birth can be easily filtered out by selections in corresponding facets.

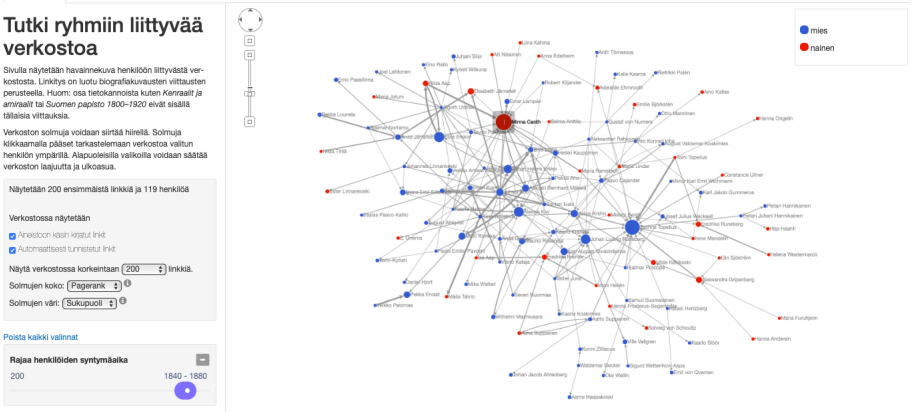


**Fig. 2.** Minna Canth in a sociocentric network where nodes are colored by gender.

An example of a group view is presented in Fig. 2. The facets are situated on the left hand side of the screen underneath the general network analysis toggles, but are not visible here. In this case, the user has filtered out Finnish authors of the mid 19th century. Here the Finnish female playwright and social activist Minna Canth gains the highest pagerank, illustrated by the size of her node. The gender of persons is indicated by red (women) or blue (men), an option selected from the toggles on the left.

**3. Explaining References** BiographySampo also contains an application view that explains the edges in the ecocentric and sociocentric networks. This reference view can be found for each protagonist in a separate tab on their homepage. The idea is to explain edges by providing the user with the sentences is which the references to other people are mentioned.[14] The sentences can be retrieved from the linguistic graph of the underlying SPARQL endpoint [35]. The references have been divided into two groups: 1) Sentences in other bios that make a reference to the protagonist's biography. 2) Sentences in the protagonist's biography that make reference to other biographies. For example, the references

[14] The view currently lists only sentences that contain manually added HTML links.

to Minna Canth include sentences from the biographies of actors, writers, and playwrights that were influenced by her, whereas her own biography mentions mainly contemporary writers and artists.

In addition to listing sentences that include links, BiographySampo also has a separate statistics application view that depicts *at what time* a person is referenced to. Here time is based on the birth year of the protagonist in the biography making the reference. The purpose of this temporal view is to be able to see how a person is referenced through time. For example, as shown in the Fig. 3, Minna Canth is frequently mentioned over a long time, because, e.g., the actors and directors of the national biography are using her plays. In comparison, a person such as the 19th century philosopher and Finnish statesman Johan Vilhelm Snellman, who had a significant role in improving the role of the Finnish language in the 19th century Finland, is mentioned, as shown in the Fig. 4, frequently mostly in the biographies of his contemporaries.

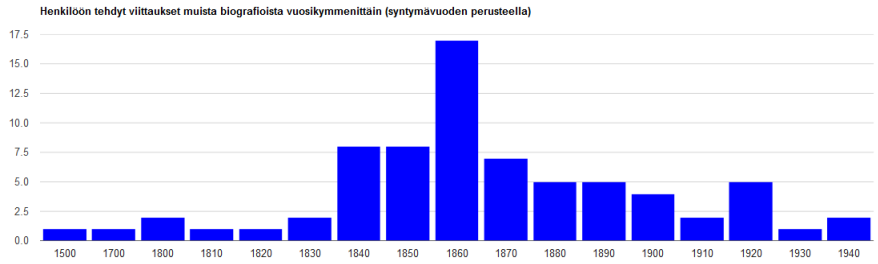


**Fig. 3.** The references made to Minna Canth.

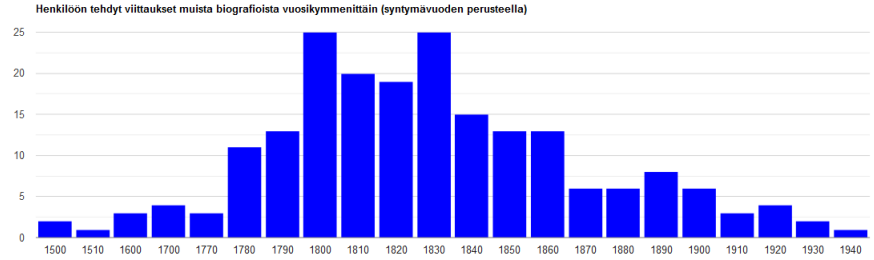


**Fig. 4.** The references made to Johan Snellman.

This view can be used to identify the influences of these notable Finns in history but also to study the edges that exist in the networks. This helps the user to see why an edge exists between two people and what kind of semantic meaning each edge holds in the network. For example, in the references page of Minna Canth, the user can see that in most cases she is mentioned because of

her literary work, and by people who have acted, directed, or visited her salon to discuss and exchange thoughts on literature and ideologies, such as Darwinism.

**4. Contextual Reader** Contextual reader is still another application of NELLI in BiographySampo. The idea here is to show the text annotated with links to the named entities, such as people, places, and organizations. It enhances the reading experience of the user by providing contextual linked information about the named entities in the biographies when the mouse is hovered over the text. The application is in work in Fig. 5, where the mouse is over *Nikolai Gogol*.

To achieve this, NELLI was configured to link named entities to contextual background information in the BSKG and other datasets available in SPARQL endpoints, such as biographies, map services, and ontology services. This was done to interlink biographical texts to each other and to help the user to understand and learn better from the texts based on their context. [26,22]

The system is based on the CORE tool [24], where entity mentions in texts can be linked to linked data resources in real time. Here string-based semantic disambiguation is used and only one interpretation is always selected. In contrast, in BiographySampo annotations are created in a pre-processing phase facilitating deeper analysis and disambiguation of entities, where challenging multiple interpretations can be given to the end-user for final human disambiguation.

This application was integrated into the biography tab of a person's homepage. The user can read the biography and gain more understanding through the links to people (indicated in blue color), places (green), and organizations (gray) (cf. Fig. 5). The links are also indicated by a symbol showing the type of the link. By hovering on top of an internal link (to BSKG) or an external link (to, e.g., Wikidata), as in Fig. 5 to the Russian playwright Nikolai Gogol, the user gets more information about that person (here from the Wikidata SPARQL endpoint). The place links lead to a map view to provide information related to that place and a map marking the location.

## 4 Assessment and Evaluation

In this section, lessons learned in developing the applications are first discussed and their usefulness assessed from an end user perspective. After this, an evaluation of the named entity linking tool NELLI follows.

**Assessing Applications** The network analysis views have been built for individuals and for groups of people. The egocentric networks for individuals are often smaller in size and therefore facets are not included in the view for filtering out related people. However, in some cases it may be interesting to scale egocentric networks to include only occupational references or people who have lived at the same time. The basic network toggling tools are provided for both views and can be used to color the nodes by occupation or gender. Also, it is possible to compare networks based on manual and automatic links.

The reference explanation view adds textual context to the links and in most cases is a helpful tool for understanding relations between nodes. However, the


**Fig. 5.** Contextual Reader application used on Kasper Järnefelt's biography.

view currently only shows the sentences with manual HTML links. In addition, in some cases one sentence does not have enough context for an explanation. For example, in the biography of *Aale Tynni* (a poet) there is a highly serendipitous surprising link to *Tapio Rautavaara* (a singer, actor, and athlete). It turns out that both of them got a gold medal in the Olympic Games in London 1948, but in different categories: Tynni in lyrics and Rautavaara in javelin throw. However, the sentence with the link does not explain this. The information is in the previous sentence, and it would be useful in this case to show more than one sentence to explain the relation.

The contextual reader application visualizes the extracted named entities in the text and adds more contexts though linking to BSKG and external datasets and ontologies. There are currently only three types of named entities visible in the contextual reader to provide context but more could be added, such as named works of art. Also, it would be useful to add a map or images for places as has been done in [29,13]. The extracted named entities from the texts are often people that the authors of the biographies consider significant occupationally. The networks and the reference analysis reflects these choices creating biases similarly to [38]. Reference networks in our case are not actual social networks. For example, in the biography of *Jutta Urpilainen*, the former Prime Minister *Jyrki Katainen* is mentioned because *Urpilainen* worked as the Minister of Finance in *Katainen's* Cabinet of Finland, but in the biography of *Katainen*, *Urpilainen* is not mentioned. It is important to keep in mind that these networks only give insight to who are considered by authors and their sources to be significant to the protagonist [15].

**Evaluation of Named Entity Extraction** In order to measure the quality of the Nelli in the task of identifying named entities, we inspected place and person links for 50 biographical texts. Self-references to the protagonist were ignored in calculations because the idea was to identify information that helps the reader to understand better about this person and the references to self

do not add value in this task. In addition, we calculated organization names containing a linked place name (e.g., The National Museum of Finland) as false positive. The linking of places and people was evaluated using precision, recall, and F1-score as shown in Table 1; the identification of organizations was ignored in the test as this is still on going work.

Table 1: Results for recognition and linking places and people.

| | Entities | $TP$ | $FP$ | $FN_{all}$ | $FN_{out}$ | $Precision$ | $Recall_{out}$ | $Recall_{all}$ | $F1_{out}$ | $F1_{all}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| **Places** | 823 | 655 | 168 | 77 | 43 | 80 % | 94 % | 89 % | 86 % | 84 % |
| **People** | 348 | 339 | 9 | 227 | 119 | 97 % | 74 % | 60 % | 84 % | 74 % |

The results in the Table 1 for places and people have been counted in two ways: 1) to exclude false negatives that cannot be found from the ontology ($FN_{out}$) and 2) to include all false negatives ($FN_{all}$). By comparing these two counts, it can be seen how entities missing from the used ontologies impacts the results for places and especially people. In most cases, the tool is dependent on the chosen ontology due to having only few people with same names. However, the overall $F1_{all}$-scores are good for people 74 % and places 84 %. The precision ($Precision$) for places is lower than the recall, causing a drop in the F1-score. In comparison, the precision for people is nearly perfect. However, the recall ($Recall_{out}$, $Recall_{all}$) for people is lower than the recall for places. This is because some people cannot be found from the ontology due to tool errors, incorrect data (missing maiden or married names, badly formed data labels), or problems with baseforming foreign names.

The precision for places suffers also due to mixing last names with place names when the names are not identified from the text. In order to reduce mixing of place and person names, the last names could be identified using the extracted full person names. Often people are referenced in the text first with a full name and later with only the last name. By using the last names from the full names, most references could be extracted and mix-ups with place names avoided. The place recognition often mixes place names and regular words, such as adjectives as places. For example, when the initial word of a sentence is the infected form of the word *oma* (engl. own), it is understood as the place Oman. By adding a rule that only considers entities that are written with a capital letter can help to reduce these issues. However, it alone is not enough and utilization of linguistic information can help to filter initial words that are not proper nouns.

## 5 Conclusions

In this case study, a total of 31 500 manually created links between biographies were utilized to visualize and study a reference network underlying a dictionary of biographies. In addition, the application of Nelli to the data added a total of 33 120 named entity links in the network of which some 12 800 were for places and some 20 800 for people. This data was utilized to enrich the networks with additional references to people cataloged in the dictionary of biography and with

new external nodes in the network. NELLI succeeded in identifying people with 74 % accuracy and places with 84 % accuracy. Four application views were added in BiographySampo to support analysis of the networks for the end user.

The selection of ontologies has a role in the success of the work. The place names in biographical texts were distinctive and easy to link to comprehensive ontologies with low granularity. It was helpful, too, that the BSKG contained only a handful of namesakes. By adding fixes to prevent the linking of adjectives and postpositions to places it is possible to increase the success rates. The disambiguation scheme used enabled successful linking of person names, which prevented most of the mix-ups between people and places.

The applications presented were based on reference networks. Unlike in [38,3,31], the user can study the networks of different groups through facet selections and visualize the networks in a variety of ways, such as re-sizing the nodes based on their topological properties or by coloring the nodes based on occupational area or gender. The networks of individuals can be studied in the egocentric network to see, for instance, the spreading of influences. The foreign influences of notable writers, politicians, and philosophers are prominent in the automatically enriched networks, and a full view of their reach can be seen through the egocentric networks. The networks, complemented with the reference view to study the explanations for the edges, gives more insight into the impact of individuals in groups. The contextual reader application enhances the reading experience by providing information about the linked entities. These applications facilitate novel, more diverse usage of BiographySampo in biographical and prosopographical research.

However, the applications also raise new questions and problems of source criticism regarding the quality of the automatically extracted content and semantic interpretation of the networks [15]. It is clear, for example, that the people selected in the dictionary not necessarily constitute a homogeneous prosopographical group but were selected by the editors, and people mentioned in the texts reflect the decisions made by the authors and the sources they have used.

**Related Work** Representing and analyzing biographical data is a new research and application field [15,36]. The network analysis based on biographical data has been studied in [38,19,3] where networks were created using a variety methods to extract named entities and their relations from text. In BiographySampo, the networks were created using the named entity linking approach [28,26,25].

The network analysis views were constructed to study individuals and groups of people. Several related works [38,31,19,3,5] and network analysis and visualization methods [27] have influenced the tools presented in this paper. The tools in BiographySampo extend traditional systems by adding user controls that can be used to scale and toggle the layout of the networks. In addition, the sociocentric network analysis allows the user to use facets (such as gender, vocation, birth and death places) to form groups of people and study their networks. To extend the network analysis tools, BiographySampo also includes a reference

analysis view explaining the links, which is similar to KORP's[15] [2] keywords in context view but provides context for the edges in the network similarly to LinkedJazz's [16] [31] relationship view. Unlike in LinkedJazz, the view shows all relations and how a person is referenced throughout time to show how a person's work influences person's contemporaries and other generations of notable people. The view is constructed using text that has been transformed into RDF [35] and by quering the sentences with manually crafted links from the SPARQL service.

In order to visualize the named entities, a contextual reader application [24] was created. Similar visualizations of named entity data have been used in, e.g., DBpedia Spotlight[17] [26] and Gate Cloud[18] [25]. The WarSampo [13] portal and the Semantic Finlex portal [29] include contextual reader applications that have been configured to link text into ontologies in real-time. These applications have influenced the creation of the BiographySampo's contextual reader. However, in our case the entities are not extracted in real time but in a preprocessing phase for more robust semantic disambiguation.

**Acknowledgments** Our research was part of the Severi project[19], funded mainly by Business Finland. Thanks to Mikko Kivelä for inspirational discussions and CSC IT Center for Science for computational resources.

## References

1. Aylett, R.S., Bental, D.S., Stewart, R., Forth, J., G.Wiggins: Supporting Serendipitous Discovery. In: Digital Futures (Third Annual Digital Economy Conference), 23-25 October, 2012, Aberdeen, UK (2012)
2. Borin, L., Forsberg, M., Roxendal, J.: Korp – the corpus infrastructure of Språkbanken. In: Proceedings of LREC 2012. Istanbul: ELRA. pp. 474–478 (2012)
3. Brouwer, J., Nijboer, H.: Golden Agents. A web of linked biographical data for the Dutch Golden Age. In: BD2017 Biographical Data in a Digital World 2017, Proceedings. vol. 2119, pp. 33–38. CEUR Workshop Proceedings (2018)
4. Bunescu, R.C., Pasca, M.: Using Encyclopedic Knowledge for Named entity Disambiguation. In: EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics. vol. 6, pp. 9–16 (2006)
5. Elson, D.K., Dames, N., McKeown, K.R.: Extracting Social Networks from Literary Fiction. In: Proceedings of the 48th annual meeting of the association for computational linguistics. pp. 138–147. Association for Computational Linguistics (2010)
6. Ferragina, P., Scaiella, U.: Tagme: on-the-fly annotation of short text fragments (by Wikipedia entities). In: Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 1625–1628. ACM (2010)
7. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R.: Evaluating entity linking with Wikipedia. Artificial Intelligence 194, 130–150 (Jan 2013)

[15] `https://korp.csc.fi/`
[16] `https://linkedjazz.org/`
[17] `https://www.dbpedia-spotlight.org/demo/`
[18] `https://cloud.gate.ac.uk/`
[19] `http://seco.cs.aalto.fi/projects/severi`

8. Hakosalo, H., Jalagin, S., Junila, M., Kurvinen, H.: Historiallinen elämä - Biografia ja historiantutkimus. Suomalaisen Kirjallisuuden Seura (SKS), Helsinki (2014)
9. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool (2011)
10. Heino, E., Tamper, M., Mäkelä, E., Leskinen, P., Ikkala, E., Tuominen, J., Koho, M., Hyvönen, E.: Named entity linking in a complex domain: Case second world war history. In: International Conference on Language, Data and Knowledge. pp. 120–133. Springer-Verlag (2017)
11. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating NLP using linked data. In: The Semantic Web – ISWC 2013. pp. 98–113. Springer Berlin Heidelberg (2013)
12. Hyvönen, E.: Publishing and using cultural heritage linked data on the semantic web. Morgan & Claypool, Palo Alto, CA (2012)
13. Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J., Mäkelä, E.: WarSampo Data Service and Semantic Portal for Publishing Linked Open Data about the Second World War History. In: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S.P., Lange, C. (eds.) The Semantic Web Latest Advances and New Domains (ESWC 2016). pp. 758–773. Springer-Verlag (2016)
14. Hyvönen, E., Ikkala, E., Tuominen, J.: Linked Data Brokering Service for Historical Places and Maps. In: Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe). vol. 1608, pp. 39–52. CEUR Workshop Proc. (2016)
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: The Semantic Web: ESWC 2019. Springer-Verlag (2019), in print
16. 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). vol. 2084, pp. 372–385. CEUR Workshop Proceedings (2018)
17. Ikkala, E., Tuominen, J., Hyvönen, E.: Contextualizing Historical Places in a Gazetteer by Using Historical Maps and Linked Data. In: Proceedings of Digital Humanities 2016, short papers. pp. 573–577 (2016)
18. Kettunen, K., Mäkelä, E., Ruokolainen, T., Kuokkala, J., Löfberg, L.: Old Content and Modern Tools-Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910. arXiv preprint arXiv:1611.02839 (2016)
19. 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 (2016)
20. Leskinen, P., Hyvönen, E.: Extracting Genealogical Networks of Linked Data from Biographical Texts (2019), `http://seco.cs.aalto.fi/publications/`, paper submitted for review
21. Leskinen, P., Hyvönen, E., Tuominen, J.: Analyzing and Visualizing Prosopographical Linked Data Based on Biographies. In: BD2017 Proceedings of the Second Conference on Biographical Data in a Digital World 2017. vol. 2119, pp. 39–44 (2018)
22. Lindquist, T., Long, H.: How can educational technology facilitate student engagement with online primary sources? A user needs assessment. Library Hi Tech 29(2), 224–241 (2011)

23. Mäkelä, E.: Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text. In: Proceedings of the ESWC 2014 demonstration track. pp. 424–428. Springer-Verlag (2014)
24. Mäkelä, E., Lindquist, T., Hyvönen, E.: CORE – a contextual reader based on linked data. In: Proceedings of Digital Humanities 2016, Krakow, Poland (long papers). pp. 267–269 (2016)
25. Maynard, D., Roberts, I., Greenwood, M.A., Rout, D., Bontcheva, K.: A framework for real-time semantic social media analysis. Journal of Web Semantics 44, 75–88 (2017)
26. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th international conference on semantic systems. pp. 1–8. ACM (2011)
27. Newman, M.: Networks. Oxford University Press (2018)
28. Nguyen, D.B., Hoffart, J., Theobald, M., Weikum, G.: AIDA-light: High-Throughput Named-Entity Disambiguation. In: Proceedings of LDOW, Linked Data on the Web. vol. 1184. CEUR Workshop Proceedings (2014)
29. Oksanen, A., Tuominen, J., Mäkelä, E., Tamper, M., Hietanen, A., Hyvönen, E.: Law and Justice as a Linked Open Data Service (2019), `http://seco.cs.aalto.fi/publications`, submitted for review
30. Otte, E., Rousseau, R.: Social network analysis: a powerful strategy, also for the information sciences. Journal of information Science 28(6), 441–453 (2002)
31. Pattuelli, M.C., Miller, M., Lange, L., Thorsen, H.K.: Linked Jazz 52nd Street: A LOD Crowdsourcing Tool to Reveal Connections among Jazz Artists. In: Proceedings of Digital Humanities 2013. pp. 337–339 (2013)
32. Piccinno, F., Ferragina, P.: From TagME to WAT: A New Entity Annotator. In: Proceedings of the first international workshop on Entity recognition & disambiguation. pp. 55–62. ACM (2014)
33. Roberts, B.: Biographical Research. Understanding social research, Open University Press (2002)
34. Small, H.: Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science 24(4), 265–269 (1973)
35. 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 (November 2018)
36. Tuominen, J., Hyvönen, E., Leskinen, P.: Bio CRM: A Data Model for Representing Biographical Data for Prosopographical Research. In: BD2017 Biographical Data in a Digital World 2017, Proceedings. vol. 2119. CEUR Workshop Proceedings (2018)
37. Verboven, K., Carlier, M., Dumolyn, J.: A short manual to the art of prosopography. In: Prosopography approaches and applications. A handbook, pp. 35–70. Unit for Prosopographical Research (Linacre College) (2007)
38. 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(3) (2016)