Using Biographical Texts as Linked Data for Prosopographical Research and Applications

Minna Tamper[1](✉), Petri Leskinen[1], Kasper Apajalahti[1],
and Eero Hyvönen[1,2]

[1] Semantic Computing Research Group (SeCo), Aalto University, Helsinki, Finland
{minna.tamper,petri.leskinen,kasper.apajalahti,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 argues that representing texts as semantic Linked Data provides a useful basis for analyzing their contents in Digital Humanities research and for Cultural Heritage application development. The idea is to transform Cultural Heritage texts into a knowledge graph and a Linked Data service that can be used flexibly in different applications via a SPARQL endpoint. The argument is discussed and evaluated in the context of biographical and prosopographical research and a case study where over 13 000 life stories form biographical collections of Biographical Centre of the Finnish Literature Society were transformed into RDF, enriched by data linking, and published in a SPARQL endpoint. Tools for biography and prosopography, data clustering, network analysis, and linguistic analysis were created with promising first results.

1 From Text to Semantic Structures

Digital Humanities (DH) [3] is a major new research paradigm at the crossroads of computing, humanities, and social sciences. The main idea is to develop and use novel computational methods, such as data analysis, topic modeling, visualization, and network analysis, to solve research problems in Social Sciences and Humanities (SSH) based on big data that is becoming available as a result of digitalization of the society.

Much of the primary data of DH is available only in textual form, and there is an ever-growing need for structuring it for semantic analysis. The research hypothesis of this paper is that representing texts as semantic Linked Data (LD), based on standards, data models, and best practices of W3C[1], provides a useful basis [6] for analyzing their contents in DH research and for Cultural Heritage (CH) application [9] development: Firstly, the data can be published

[1] https://www.w3.org/standards/semanticweb/ accessed: 13 August 2018.

M. Ioannides et al. (Eds.): EuroMed 2018, LNCS 11196, pp. 125–137, 2018.
https://doi.org/10.1007/978-3-030-01762-0_11

in standard RDF formats in a data service on the Semantic Web that can be queried in flexible ways for extracting the data for different use cases. Secondly, the Linked Data paradigm facilitates data enrichment by data linking and fusion from related data repositories. Thirdly, the semantics of LD is defined in terms of logic, which facilitates data enrichment by reasoning [8].

This paper investigates and evaluates this hypothesis by four case studies. First (Sect. 2) a natural language (NL) pipeline for transforming texts into a knowledge graph to be published in a SPARQL endpoint service is presented. After this (Sect. 3), we investigate using the graph in four different use cases in order to test and demonstrate the versatility of the approach. As for the data, the National Biography of Finland, a collection of short textual biographies in addition to other peer-reviewed biographical collections from the Biographical Centre of the Finnish Literature Society totaling in 13 000 life stories, are used. In conclusion (Sect. 4), contributions of the paper and lessons learned are summarized, and related work discussed. The novelty of this paper regarding our earlier publications [13,17] about the Semantic National Biography of Finland (SNBF), is to present the underlying NL transformation pipeline in detail and to show how new tools for prosopography related to clustering, network analysis, and linguistic analysis, can be added on top of the LD service.

2 A Pipeline for Transforming Text into Linked Data

For transforming Finnish texts into knowledge graphs, we have created a general NL pipeline. The pipeline has two branches: (1) one for semi-structured text and (2) one for free text. This is because a part of the texts in focus in our use case, i.e., in biographies, are written in a concise, semi-formal way, explicating the major events, achievements, and other biographical data about the protagonist [28]. Here, for example, listings and abbreviations are widely used (for educational degrees, professions, honorary medals, etc.), and verbs are rarely used for brevity.[2] The main life story, on the other hand, is written in terms of normal full sentences.

The target data model in our study is Bio CRM, an extension of the CIDOC CRM ISO standard[3] for biographical data. The key idea of the model is to represent biographies as sequences of events that the protagonist participated in space and time and in different roles. The Bio CRM model is presented in more detail in [26].

Pattern-Based Knowledge Extraction. For the semi-structured part of the bios, extraction rules based of regular expressions were used. This part includes, for example, descriptions of family relations of the protagonist and lists of her/his

[2] In some use cases, e.g., in person registries [12], the whole registry entry may be written using this kind of semi-formal language.

[3] http://cidoc-crm.org accessed: 13 August 2018.

professional history. An example of such descriptions for the architect *Eliel Saarinen* is given below:

> Gottlieb Eliel Saarinen S 20.8.1873 Rantasalmi, K 1.7.1950 Bloomfield Hills, Michigan, Yhdysvallat. V rovasti Juho Saarinen ja Selma Maria Broms. P1 1898 - 1902 (ero) Mathilda Tony Charlotta Gylden (sittemmin Gesellius) S 1877, K 1921, P1 V agronomi Axel Gylden ja Antonia Sofia Hausen; P2 1904 - kuvanveistäjä Minna Carolina Louise (Loja) Gesellius S 1879, ...

The semi-formal expressions here have uniformity in structure that can be used effectively for pattern-based information extraction: First, the person's given and family names are mentioned and after that the fields of birth and death information are separated with *S* for birth, and *K* for death. These fields contain the time and place of the event. A field beginning with *V* contains the information about the person's parents with father followed by mother, their names, occupations, and possible places and times of birth and death. Likewise, fields beginning with *P*, or if several *P1*, *P2* etc., carry the information of possible spouses indicating the year of marriage, and the spouse's living time. The data field may also contain information about the parents of the spouse in *PV*, *PV1*, *PV2*, etc. fields. At the end of the description there is a list of children with their names, occupation, and times and places of birth and death. The process has running time complexity *O(n)* in relation to the amount of text descriptions, and extracting all the 52 476 family relations from the dataset took 5.2 s.

The semi-structured text also includes the events of the person's career:

> URA. Käynyt kaksi luokkaa Viipurin suomalaista klassillista lyseota 1883 - 1887, kolme luokkaa Viipurin Alkeiskoulua 1887 - 1890; ylioppilas Tampereen reaalilyseosta 1893; arkkitehti Suomen Polyteknillisestä opistosta 1897; ...

Here Eliel Saarinen's career events, beginning with the word "URA", are separated into parts with semicolons. Each part contains text indicating, e.g., the person's education, employer, or other description, followed by years. Entities and their relations in the semi-structured texts were extracted with Python scripts using regular expressions. From this data altogether 102 004 life time Bio CRM events were generated. Extracting the events took 147.6 s and they used 330.0 Mb of disk space.

Knowledge Extraction from Free Text. For the free text part, more complex NL processing is needed. Like in NewsReader [24] and FRED [21], this pipeline branch was built using pre-existing NLP tools. The process consists of linguistic analyses (such as tokenization and morphological tagging) and converting the document structures and the linguistic data into RDF. The NLP Interchange Format (NIF)[4] [7] supplements the RDF representation with a Core Ontology that provides classes and properties to describe the relations between texts and documents. This provides flexibility and structure to divide a document into paragraphs, titles, sentences, and words that can be complemented with structural metadata supplied by NIF and linguistic information, such as lemmas and part-of-speech (POS) tags from NLP tools. In addition to the NIF

[4] http://persistence.uni-leipzig.org/nlp2rdf/specification/core.html accessed: 13 August 2018.

format, the commonly used CIDOC CRM ISO standard, Dublin Core Metadata[5], and custom namespace[6] are used to supply classes and properties for describing document metadata.

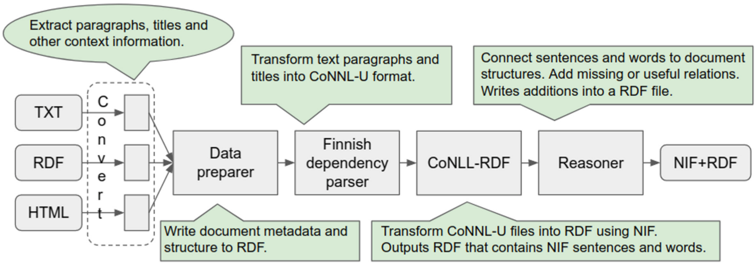

Extract paragraphs, titles and other context information.
TXT
RDF
HTML
Convert
Data preparer
Finnish dependency parser
CoNLL-RDF
Reasoner
NIF+RDF
Transform text paragraphs and titles into CoNNL-U format.
Connect sentences and words to document structures. Add missing or useful relations. Writes additions into a RDF file.
Write document metadata and structure to RDF.
Transform CoNNL-U files into RDF using NIF. Outputs RDF that contains NIF sentences and words.

Fig. 1. Pipeline for text processing

The model for transforming text into RDF can be seen in Fig. 1. The pipeline supports RDF, HTML, and text input formats. Their processing starts by extracting paragraphs and titles from each document. The use of RDF input format requires that there is a pre-existing document structure in RDF that can be mapped to the NIF format by the converter. The HTML documents are split into paragraphs utilizing `p` tags whereas the text document is split to paragraphs based on the assumption that there is an empty line separating the paragraphs from one another. The titles are picked from HTML using the `h` tags for headers of different levels. From the text input, titles are picked using regular expressions with the assumption that a title is a paragraph that never ends with a dot. After the extraction and conversion phase, the Data preparer writes an RDF file that describes the document structures using CIDOC CRM class *crm:E31_Document* for documents, and NIF format classes *nif:Title* and *nif:Paragraph* for titles and paragraphs within the text document to record substrings using *nif:isString* property and to connect the substrings with *dct:isPartOf* to the document instance representing the full text. These structures are accompanied by properties to describe document sources (using *nbf:docRef* to store the identifier to the original document), substring order, and other structural data, such as HTML links included in the original text. The HTML links are currently added to the *nif:Paragraph* instances using *dct:references* and defined as instances of custom *nbf:Anchor* class that has *nif:isString* property for the anchor texts and *nbf:anchor_link* for the links. The Data preparer module does not structure text but outputs the substrings in separate text files for the following phases.

After the Data preparer phase, the pipeline proceeds to execute the Finnish dependency parser[7] [5] as shown in Fig. 1. Using the texts produced by the

[5] http://dublincore.org/documents/dcmi-terms/ accessed: 13 August 2018.
[6] Denoted with prefix nbf in the RDF examples.
[7] http://turkunlp.github.io/Finnish-dep-parser/ accessed: 13 August 2018.

data preparer, the parser transforms the texts into CoNNL-U[8] [22] format for each paragraph and title. For example, from the biography of the architect *Eliel Saarinen*, the title and paragraphs are analyzed and the application produces separately for each of them a file containing sentences and words, their positions, lemmas, POS tags, morphological features, dependencies, and other linguistic information. The results of the transformation of text to CoNLL-U format produces a file where each word or token of a sentence is represented on a separate line with original form, lemma, and linguistic information. The sentences are separated with empty lines. The Finnish dependency parser was selected for this task because it is an open source tool, easy to plug in, and reliable (estimated accuracy is 81% [5]) for Finnish language texts. However, if the tool's performance does not yield satisfactory results, it can be complimented[9] or replaced with other tools.

The transformation of text to CoNNL-U format is followed by conversion into RDF using the CoNLL-RDF tool [1] that transforms sentences and words into RDF using corresponding *nif:Sentence* and *nif:Word* classes. In addition, the CoNLL-RDF tool generates the identifiers for the instances of *nif:Sentence* and *nif:Word* classes and adds the CoNNL-U data as properties to provide linguistic information of the words. The tool writes the data into an RDF file. For example, the title of the biography of Eliel Saarinen is split to several tokens that form one sentence. The output of this tool can be shown in the Fig. 2 where the title of the protagonist's bio (*Saarinen, Eliel*) is presented in RDF format.

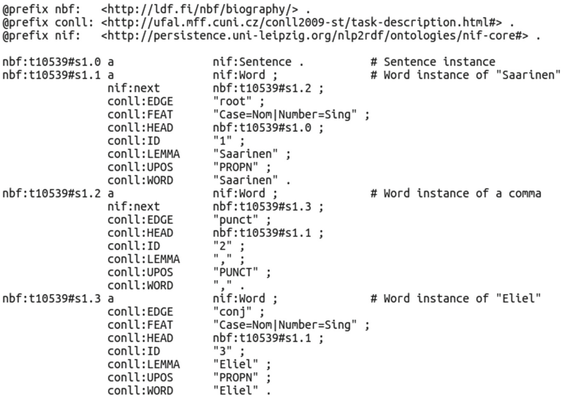
The results of the CoNLL-RDF transformation are next used in the Reasoner module (cf. Fig. 1) that uses RDF files of the previous phase to add missing relations between the sentences, words, and the general document structure (namely paragraphs, titles and biography documents). The results of the reasoner are shown in Fig. 3. The reasoner has added *dct:isPartOf* relation for the *nif:Sentence* instance to attach it to the *nif:Title* and *crm:E31_Document* instances. In addition, the sentence instance is complemented with an order number that has been deduced from the sentence identifier shown in Fig. 2. Similarly, the instances of the *nif:Word* have been supplied with *dct:isPartOf* and *nif:sentence* relations to connect it with the biography and the sentence correspondingly. Similarly to the sentence order property, the sentence identifier is deduced from the word instance's identifier. This enables, for example, the user to query words relating to a particular biography document, paragraph, or sentence (this is needed in linguistic analyses). In addition, by adding the order property for sentences, it is easy to query the sentences in correct order. Lastly, the Reasoner component writes an RDF file that contains all of these relations. Once the text is processed, the pipeline returns a set of RDF files or can upload them directly into a SPARQL endpoint. The process was executed for all life stories (total of 13 000 biographies). It runs in linear time[10], and with these life

[8] http://universaldependencies.org/format.html accessed: 13 August 2018.

[9] For example, the SeCo LAS [19] is a combination of several Finnish NLP tools.

[10] The running time complexity *O(mn)*, where n is the amount of files and m their size in bytes.

stories it takes on average 1.9 s for each file (totaling in 161 076 files containing titles and paragraphs using 7.4 GB of disk space).



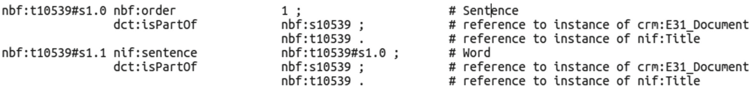
```
@prefix nbf:   <http://ldf.fi/nbf/biography/> .
@prefix conll: <http://ufal.mff.cuni.cz/conll2009-st/task-description.html#> .
@prefix nif:   <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .

nbf:t10539#s1.0 a               nif:Sentence .          # Sentence instance
nbf:t10539#s1.1 a               nif:Word ;              # Word instance of "Saarinen"
                nif:next        nbf:t10539#s1.2 ;
                conll:EDGE      "root" ;
                conll:FEAT      "Case=Nom|Number=Sing" ;
                conll:HEAD      nbf:t10539#s1.0 ;
                conll:ID        "1" ;
                conll:LEMMA     "Saarinen" ;
                conll:UPOS      "PROPN" ;
                conll:WORD      "Saarinen" .
nbf:t10539#s1.2 a               nif:Word ;              # Word instance of a comma
                nif:next        nbf:t10539#s1.3 ;
                conll:EDGE      "punct" ;
                conll:HEAD      nbf:t10539#s1.1 ;
                conll:ID        "2" ;
                conll:LEMMA     "," ;
                conll:UPOS      "PUNCT" ;
                conll:WORD      "," .
nbf:t10539#s1.3 a               nif:Word ;              # Word instance of "Eliel"
                conll:EDGE      "conj" ;
                conll:FEAT      "Case=Nom|Number=Sing" ;
                conll:HEAD      nbf:t10539#s1.1 ;
                conll:ID        "3" ;
                conll:LEMMA     "Eliel" ;
                conll:UPOS      "PROPN" ;
                conll:WORD      "Eliel" .
```

Fig. 2. Output of the CoNLL-RDF tool

3 Analyzing and Using Biographies as Knowledge Graphs

This section presents how the pipeline of Sect. 2 was used in four use cases, illustrating and evaluating the research hypotheses presented in Sect. 1.



```
nbf:t10539#s1.0 nbf:order               1 ;                     # Sentence
                dct:isPartOf            nbf:s10539 ;            # reference to instance of crm:E31_Document
                                        nbf:t10539 .            # reference to instance of nif:Title
nbf:t10539#s1.1 nif:sentence            nbf:t10539#s1.0 ;       # Word
                dct:isPartOf            nbf:s10539 ;            # reference to instance of crm:E31_Document
                                        nbf:t10539 .            # reference to instance of nif:Title
```

Fig. 3. Output of the reasoner

Case Study Setting and Data. The National Biography of Finland[11] (NBF), edited by the Finnish Literature Society, is a collection of biographies written by the experts in the fields of history, science, art, culture, and business. The collection consists of several datasets: National Biography core, Admirals and Generals of Finland, Finnish Clergy, and Business Leaders. There are altogether 13 000 biographies [13]. The data was available in CSV format that contains for each person a biographical description article, basic information, such person name and lifetime information, and article level metadata, such as the author and date of publishing. The article about the protagonists written in free text followed by semi-formal synopsis text, as presented above, In addition, the HTML versions of the biographies contain manually annotated internal links to other

[11] https://kansallisbiografia.fi/ accessed: 13 August 2018.

biographies. They have been constructed by the editors and include the person's name as an anchor text and a link to the person's biography.

The biographies were transformed into RDF by using the pipelines of Sect. 2, and were uploaded into a data service and SPARQL endpoint at the Linked Data Finland service[12] [14]. From there, the data can be easily accessed and used for biographical and prosopographical research and data analysis. For this purpose, a semantic portal was created. The portal and the underlying linked data service are called jointly "Semantic National Biography of Finland" (SNBF).

Use Case 1: Biography and Prosopography. NBF is the authoritative biography collection of notable Finns. It is used for studying individual life stories in biography research [23] by close reading. Based on transforming the bios into linked data and the data service, also prosopographical research based on distant reading [25] becomes possible. Prosopography [4,27] 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. However, it is slow and infeasible to go through thousands of biographies by hand and analyze them from different perspectives. SNBF helps here by supporting data analysis.

This use case from the end user's view point has been reported in more detail in [13,17], and we only summarize the results here. The RDF used was extracted from the semi-formal parts of the biographies. On the portal side, a faceted search tool was created and used to (1) search biographies of interest and (2) to filter out target groups for prosopographical research. The facets were based on biographical basic attributes, like time and place of birth and death, profession, name, or gender.

For biography research, "home pages" for the protagonists were created and the data was enriched by linking them to ten external data repositories and data services, such as Wikidata. The events extracted from the life stories were visualized on timelines and on maps for a spatiotemporal biographical perspective.

As for prosopography, visualizations, such as pie charts, histograms, and sankey diagrams were used for studying the properties of the target group filtered out by selections on the facets. For example, a statistic page of the group contains five column histograms illustrating people's life span, the age of marriage, the age of getting the first child, the amount of children, and the number of spouses.

The data service uses two SPARQL endpoints, the service for person ontologies at http://ldf.fi/nbf/sparql has 4 600 000 triples, and the service for linguistic data produced by the pipeline at http://ldf.fi/nbf-nlp/sparql has 120 000 000 triples (7.4 Gb of data). Our dataset includes 98 953 person entries of which 54 409 are relatives whose information was extracted from the structured textual descriptions (In addition to 13 000 protagonists there are 30 000 people mentioned in the table of contents, and 1000 mentioned as authors). In the CIDOC CRM based Bio CRM data model births and deaths are modeled as events linked to the actor. The family relations are modeled according to the Bio CRM model as relations with corresponding roles [26]. The network of family relatives was build using only simple, direct relations like Parent with

[12] http://www.ldf.fi accessed: 13 August 2018.

subclasses Mother and Father, Child with subclasses Daughter and Son, and Spouse. The inverse family relations were later inferred from the data using reasoning. The family relation descriptions were converted from source data into RDF using Python scripts which, e.g., extracted the time spans, and separated a person text field into his/her occupation, given, and family names by using regular expressions. Events of birth and death contain the time and the place name. The place names were linked to the place ontology of SNBF [17] using SPARQL ARPA tool[13]. The place ontology consists of Finnish places extracted from Hipla [11,15], and of Foreign places located with Google Maps API[14]. Altogether the process produced 54 409 people, 23 762 births, 15 952 deaths, and 88 356 family relations.

Use Case 2: Clustering Data for Recommending. The research problem addressed in this use case is: Given a person in the biography collection, are there other people with "similar" lives, and how could these people be found automatically? Finding out such clusters of similar people could give insight on what kind of groups there actually are in the biography collection. A user reading a biography is likely interested in reading more biographies of similar kind, so clustering could also be used as a basis for a recommender system [16]. A critical question here is what criteria for similarity to use since there are lots of options available.

To test and demonstrate the potential of clustering in recommending, a recommender system was implemented in SNBF. A similarity between two people was defined as the cosine similarity between the TF-IDF[15] vectors of their biographies. TF-IDF is one of the most fundamental methods used in information retrieval, and gives a kind base line for testing clustering.

To calculate the distance mapping, we first made a SPARQL query collecting all the nouns, adjectives, and verbs in the lemmatized form. The TF-IDF embeddings were generated from these texts. The similarity map was constructed by adding a link between a pair of people if the cosine similarity between their TF-IDF vectors exceeded a pre-defined threshold value. The recommending system on the project page queries for and shows the candidates in ascending order by distance. The recommender system was implemented using the Gensim Library[16] for Python.

The results of clustering was evaluated by manually checking the linkage of a small set of people with first encouraging results. The system is capable of clustering people who share common characteristics, e.g., it forms clusters of politicians, architects, military personnel, etc. On the other hand, in cases of people that have very short bios, or when they do not belong to any specific larger group, the concluded recommended links were not so obvious.

Use Case 3: Network Analysis. Network analysis [20] can be used to study connections of individuals within a specific community. The biographical data

[13] http://seco.cs.aalto.fi/projects/dcert/ accessed: 13 August 2018.

[14] http://developers.google.com/maps/ accessed: 13 August 2018.

[15] http://www.tfidf.com/ accessed: 13 August 2018.

[16] https://radimrehurek.com/gensim/ accessed: 13 August 2018.

can be used to construct social networks to analyze the patterns of relationships, composition and activities of people in their own historical context. For this purpose, the SPARQL endpoint can be utilized to first select a target group of people with respect to desired criteria, such as time of birth, gender, or profession. From the query results, it is easy to construct social networks based on different criteria, such as family or business relations, references or co-location in biography texts, and so on. Social networks can finally be analyzed using standard network metrics and visualization tools.

For example, the SPARQL query of Fig. 4 finds people born in the 20th century and their professions. The query results can be used to construct a network of people based on the HTML links between the biographies, showing connections between the persons. A network of (cf. Fig. 5) is then created and shown using Gephi[17], illustrating the clusters of professions that are shown in different colors. The modularity, degrees, PageRank, HITS, and different centrality measures [18] can be calculated from the network to find our central figures of the community and their roles.

In Fig. 5, we have a network of people constructed using the query in Fig. 4. In this network people are divided into clusters by their professions. The black dots represent professions that have not been included in the list of professions in the query. Other colors represent professions that are described in the picture. The node size is based on the PageRank measure estimating the centrality of a node. The highest PageRank values indicate that the person references and is frequently referenced in other biographical texts. The clustering in the network indicates that the biographies make reference to people of the same vocation. The politicians, rulers, and presidents are highly clustered among themselves whereas the musicians are split to two groups representing classical and popular music.

Use Case 4: Linguistic Analysis. The linguistic data created by the pipeline can be used to study not only the protagonists and groups of them but also how the biographies have been written. When historians write texts, they arrange, interpret, and generalize from facts and events using their own style and words. By assessing biographies by looking at the use of vocabulary it is possible to (1) find and analyze differences between authors as well as (2) differences between groups of biographies written by different authors. For example, what kind of differences in wordings are there in use when describing women and men, or politicians in different parties?

To facilitate linguistic analysis and comparison of bios, a separate application view was implemented in the SNBF portal. This view utilizes the data produced by the NL pipeline. A faceted search interface, using facets such as authors, time periods, vocations, titles, genders, places, and last names, is used for focusing the analysis on different target groups of bios. By selecting facet categories, the underlying SPARQL query is modified and the application renders the query result as a linguistic analysis of the target group bios. The analysis is represented as a table that lists the verbs, common nouns, adjectives, and proper nouns

[17] https://gephi.org/ accessed: 13 August 2018.

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX categories: <http://ldf.fi/nbf/categories/>
PREFIX gvp: <http://vocab.getty.edu/ontology#>
PREFIX crm: <http://www.cidoc-crm.org/cidoc-crm/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX sources: <http://ldf.fi/nbf/sources/>
PREFIX nbf: <http://ldf.fi/nbf/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skosxl: <http://www.w3.org/2008/05/skos-xl#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT ?from ?to ?weight WHERE {
  BIND(1 as ?weight) {
    SELECT distinct ?from WHERE {
      { ?from a nbf:PersonConcept . }
      ?from foaf:focus/^crm:P98_brought_into_life/nbf:time/gvp:estStart ?birth .
      FILTER (?birth >=”1900-01-01”^^xsd:date)
      ?from dcterms:source sources:source1 .
      ?from foaf:focus/nbf:has_category ?category .
      FILTER (?category IN (categories:c133, categories:c44, categories:c41, categories:c46, categories:c131,
categories:c61,  categories:c51, categories:c43, categories:c12) )
    } ORDER BY DESC(?birth)
  }
  SERVICE <http://ldf.fi/nbf-nlp/sparql> {
    ?structure <http://ldf.fi/nbf/biography/data#docRef> ?from .
    ?paragraph dcterms:isPartOf ?structure .
    ?paragraph dcterms:referenced ?link .
    ?link <http://ldf.fi/nbf/biography/data#anchor_link> ?target_link .
  }
  ?to nbf:formatted_link ?target_link .
  ?to skosxl:prefLabel/skos:prefLabel ?label .
  ?to foaf:focus/^crm:P98_brought_into_life/nbf:time/gvp:estStart ?birth2 .
  FILTER (?birth2 >=”1900-01-01”^^xsd:date)
}
```

Fig. 4. SPARQL query for constructing a network of people

used in the bios with their frequencies. Moreover, two independent separate faceted search views are shown in parallel in the application, so that the user can compare the language used in two target groups. For example, it is possible to compare how (1) writers vs. (2) artists in the first half of the 19th century are described linguistically. The list of adjectives for writers has words, such as “scientific”, “political”, and references to different languages. For artists the most common adjectives are related to nationalities, countries, and artistic styles, such as “romantic” and “realistic”.

Through these facets the user can also get general statistics such as the number of documents by period of time, or the amount of words in these documents to understand better the division and length of the documents.

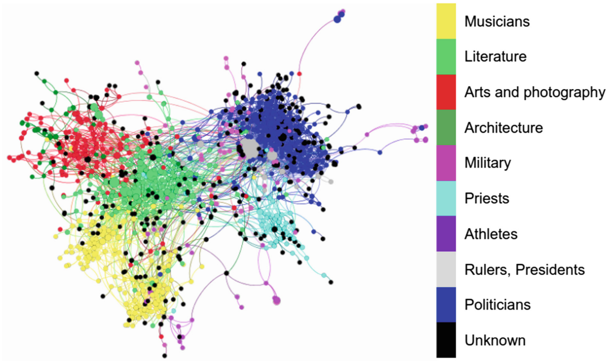


Fig. 5. Network analysis of historical people clustered by their profession

In addition to word frequency analyses, there are also other ways for analyzing the bios linguistically. For example, tag clouds for content summarization, concordance analysis of word contexts, and topic modeling with visualizations could be incorporated in the linguistic analysis view application on top of the SPARQL endpoint.

4 Discussion

Contributions and Related Work. NL pipelines for transforming text into knowledge graphs have been created, e.g., in FRED [21], NewsReader [24], and BiographyNet [2]. The pipeline presented in this paper is the first one for Finnish. A distinctive feature of the pipeline is that the whole text, word by word, is transformed into RDF form, retaining also full linguistic information of the texts. We envision that based on such a rich representation, higher level systems for knowledge extraction, named entity recognition with semantic disambiguation, relation extraction, and event extraction, can be implemented more easily, and interesting linguistic research question can be answered, too, as exemplified in our last use case. Yet another contribution of this paper is to present novel use cases for using textual biography collections as Linked Data, supported by working demonstrators. Especially, the paper discussed (1) utilizing faceted search and browsing, combined with data analysis and visualizations, for biography and prosopography, (2) biography clustering for knowledge discovery and recommender systems, (3) the network analysis, and (4) the linguistic analysis of biographical texts.

Each of the use cases opens new wide avenues for studying biographical dictionaries and collections as data, providing new insights into history and cultural heritage research. There are lots of research publications related to faceted search and browsing, data analysis, visualization, prosopography, clustering, network analysis, and linguistic analysis. Instead of discussing these fields of research in detail, our main point was to show, with the support of actual implementations, that by transforming biographies into Linked Data and by publishing the data as a Linked Data service, these technologies and software available can be combined and reused in the new setting of biography and prosopography research in promising ways. The proposed ideas can be applied not only in SNBF but also in other national and other biographical dictionaries, such as the Oxford Dictionary of National Biography[18], USA's American National Biography[19], Germany's Neue Deutsche Biographie[20], Dictionary of Swedish National Biography[21], Biography Portal of the Netherlands[22], and BiographyNet[23].

Future Work. In the future, we plan to publish more detailed accounts of the use cases, presented only shortly in this paper, in more focused research

[18] http://www.oxforddnb.com/ accessed: 13 August 2018.

[19] http://www.anb.org/ accessed: 13 August 2018.

[20] http://www.ndb.badw-muenchen.de accessed: 13 August 2018.

[21] https://sok.riksarkivet.se/Sbl/Start.aspx accessed: 13 August 2018.

[22] http://www.biografischportaal.nl/en accessed: 13 August 2018.

[23] http://www.biographynet.nl/ accessed: 13 August 2018.

papers. The Semantic National Biography of Finland will be launched as an open national on-line data service and a semantic portal in September 2018, completing the work that started in 2013 with a first spatio-temporal demonstrator [10] for biographies based on events. We will continue to work and upgrade on methods described above after the publication.

Acknowledgements. Our research is part of the Severi project (http://seco.cs.aalto.fi/projects/severi accessed: 13 August 2018), funded mainly by Business Finland.

References

1. Chiarcos, C., Fäth, C.: CoNLL-RDF: linked corpora done in an NLP-friendly way. In: Gracia, J., Bond, F., McCrae, J.P., Buitelaar, P., Chiarcos, C., Hellmann, S. (eds.) LDK 2017. LNCS (LNAI), vol. 10318, pp. 74–88. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59888-8_6
2. Fokkens, A., et al.: Biographynet: extracting relations between people and events. In: Europa baut auf Biographien, pp. 193–224. New Academic Press, Wien (2017)
3. Gardiner, E., Musto, R.G.: The Digital Humanities: A Primer for Students and Scholars. Cambridge University Press, Cambridge (2015)
4. Hakosalo, H., Jalagin, S., Junila, M., Kurvinen, H.: Historiallinen elämä - Biografia ja historiantutkimus. Suomalaisen Kirjallisuuden Seura (SKS) (2014)
5. Haverinen, K., et al.: Building the essential resources for Finnish: the Turku Dependency Treebank. Lang. Resour. Eval. **48**, 493–531 (2014). https://doi.org/10.1007/s10579-013-9244-1. Open access
6. Heath, T., Bizer, C.: Linked data: evolving the web into a global data space. Synthesis Lectures on the Semantic Web: Theory and Technology, 1 edn. Morgan & Claypool, Palo Alto (2011). http://linkeddatabook.com/editions/1.0/. Accessed 13 Aug 2018
7. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating NLP using linked data. In: Alani, H. (ed.) ISWC 2013. LNCS, vol. 8219, pp. 98–113. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41338-4_7
8. Hitzler, P., Krötzsch, M., Rudolph, S.: Foundations of Semantic Web Technologies. Springer, Heidelberg (2010)
9. Hyvönen, E.: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, Palo Alto (2012)
10. Hyvönen, E., Alonen, M., Ikkala, E., Mäkelä, E.: Life stories as event-based linked data: case semantic national biography. In: Proceedings of ISWC 2014 Posters & Demonstrations Track. CEUR Workshop Proceedings, October 2014. http://ceur-ws.org/Vol-1272/. Accessed 13 Aug 2018
11. 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 Proceedings (2016). http://ceur-ws.org/Vol-1608/#paper-06. Accessed 13 Aug 2018
12. Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., Sirola, L.: Reassembling and enriching the life stories in printed biographical registers: norssi high school alumni on the semantic web. In: Gracia, J., Bond, F., McCrae, J.P., Buitelaar, P., Chiarcos, C., Hellmann, S. (eds.) LDK 2017. LNCS (LNAI), vol. 10318, pp. 113–119. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59888-8_9

13. 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, March 2018. http://www.ceur-ws.org/Vol-2084/short12.pdf. Accessed 13 Aug 2018
14. 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
15. 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)
16. Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender Systems: An Introduction. Cambridge University Press, Cambridge (2011)
17. Leskinen, P., Hyvönen, E., Tuominen, J.: Analyzing and visualizing prosopographical linked data based on short biographies. In: Biographical Data in a Digital World 2017 (BD 2017), Linz, Austria, November 2017. http://ceur-ws.org/Vol-2119/paper7.pdf. Accessed 13 Aug 2018
18. McSweeney, P.J.: Gephi network statistics. Google Summer Code, pp. 1–8 (2009)
19. Mäkelä, E.: LAS: an integrated language analysis tool for multiple languages. J. Open Source Softw. **1**(6) (2016). https://doi.org/10.21105/joss.00035. Accessed 13 Aug 2018
20. Otte, E., Rousseau, R.: Social network analysis: a powerful strategy, also for the information sciences. J. Inf. Sci. **28**(6), 441–453 (2002)
21. Presutti, V., Draicchio, F., Gangemi, A.: Knowledge extraction based on discourse representation theory and linguistic frames. In: ten Teije, A. (ed.) EKAW 2012. LNCS (LNAI), vol. 7603, pp. 114–129. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33876-2_12
22. Pyysalo, S., Ginter, F.: Collaborative development of annotation guidelines with application to universal dependencies. In: The Fifth Swedish Language Technology Conference (2014)
23. Roberts, B.: Biographical Research. Understanding Social Research. Open University Press (2002)
24. Rospocher, M., et al.: Building event-centric knowledge graphs from news. Web Semant. Sci., Serv. Agents World Wide Web **37**, 132–151 (2016)
25. Shultz, K.: What is distant reading? New York Times, 24 June 2011. https://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html. Accessed 13 Aug 2018
26. Tuominen, J., Hyvönen, E., Leskinen, P.: Bio CRM: a data model for representing biographical data for prosopographical research. In: Proceedings of the Biographical Data in a Digital World 2017 (BD2017). CEUR Workshop Proceedings (2018). http://ceur-ws.org/Vol-2119/paper10.pdf. Accessed 13 Aug 2018
27. 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)
28. Wu, Y., Sun, H., Yan, C.: An event timeline extraction method based on news corpus. In: 2017 IEEE 2nd International Conference on Big Data Analysis, pp. 697–702. IEEE (2017)