

Building Lightweight Ontologies for Faceted Search with Named Entity Recognition: Case WarMemoirSampo

Mikko Koho¹, Rafael Leal¹, Esko Ikkala¹, Minna Tamper^{1,2}, Heikki Rantala¹ and Eero Hyvönen^{1,2}

¹*Semantic Computing Research Group (SeCo), Aalto University, Finland*

²*Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland*

Abstract

This paper discusses building lightweight ontologies for faceted search user interfaces with Named Entity Recognition (NER) from textual data. This is studied in the context of building a Knowledge Graph for the textual indexing of interview videos in the in-use WarMemoirSampo system, consisting of a Linked Open Data service and an open semantic web portal for contextualized video viewing. It is shown that state-of-the-art NER tools are able to find entities from textual data and categorize them with high enough recall and precision to be useful for building facet ontologies, without involving considerable manual domain ontology engineering. To enable entity disambiguation and to be able to show relevant contextual information and useful links for the users of the portal, also Named Entity Linking techniques are employed.

Keywords

Named Entity Recognition, Information Extraction, Faceted Search, Linked Data, Ontologies, Named Entity Linking

1. Introduction


In the 1930s, S. R. Ranganathan introduced the idea of faceted classification in Library Science. Related to this idea, the *faceted search* paradigm [1, 2], called also *view-based search* [3] and *dynamic taxonomies* [4], is based on indexing data items along orthogonal category hierarchies, i.e., facets (e.g., subject matter, places, times, etc.). The user can select categories on the facets in free order, and the data items included in the selected categories are considered the search results. After a selection, a count is calculated for each category, showing the number of results for that selection, which is useful for guiding the search. In [5], ontologies that interlink the data in a semantic web application were used as a basis for associating data to the facets. The paradigm has been found suitable for, e.g., Semantic Web user interfaces in cultural heritage

Text2KG 2022: International Workshop on Knowledge Graph Generation from Text, Co-located with the ESWC 2022, May 2022-05-30, Crete, Hersonissos, Greece

✉ mikko.koho@aalto.fi (M. Koho); rafael.leal@aalto.fi (R. Leal); esko.ikkala@aalto.fi (E. Ikkala); minna.tamper@aalto.fi (M. Tamper); heikki.rantala@aalto.fi (H. Rantala); eero.hyvonen@aalto.fi (E. Hyvönen)
🆔 0000-0002-7373-9338 (M. Koho); 0000-0001-7266-2036 (R. Leal); 0000-0002-9571-7260 (E. Ikkala); 0000-0002-3301-1705 (M. Tamper); 0000-0002-4716-6564 (H. Rantala); 0000-0003-1695-5840 (E. Hyvönen)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

(CH) applications, such as [6, 7, 8], and various tools for faceted Linked Data-based search and browsing [9] have been developed, such as */facet* [10], *SPARQL Faceter* [11], and *Sampo-UI* [12].

In the Linked Data [13] context, applications typically use pre-defined ontologies [14] for facets, the same that have been used for annotating the underlying metadata. For example, in [8] the vocabularies¹ of the Getty Research Institute were employed. Ontologies can also be created in many ways, typically involving domain experts in manually engineering the ontologies for some specific needs [15]. In many cases such vocabularies or ontologies may not be readily available, and manually creating one might not be feasible, for example, when the metadata to be searched has not been structured (like is the case in, e.g., museum collections) but are available only in textual form. The vocabularies then have to be constructed bottom-up from the textual data.

In this paper, we argue that it is possible to build *lightweight ontologies* [16, 17] by applying state-of-the-art *Named Entity Recognition (NER)* [18] methods to textual data to find named entities of different categories, and that these lightweight ontologies can be used to enable the search, exploration, and analysis of data on a faceted search user interface. We show how lightweight facet ontologies can be built from the entities found with NER and the textual objects linked to the created ontologies to enable ontology-based information retrieval.

These topics are studied in the context of WARMEMOIRSAMPO, which is a *Linked Open Data (LOD)* resource of Finnish Second World War (WW2) veteran interview videos [19], as well as an in-use semantic portal² for easy access to them. It hosts a collection of 159 videos, with approximately 400 hours of playtime in total, where veterans mostly reminisce about their lives during and after the wartime. WARMEMOIRSAMPO is realized by combining NER and *Named Entity Linking (NEL)* [20, 21, 22] on the video content descriptions, and enriching video metadata with related information from the linked entities in the WarSampo knowledge graph [23] and Wikidata [24]. The original video content descriptions consist of free-form notes in Finnish, which are time-indexed to enable finding video segments based on the discussed topics. In WARMEMOIRSAMPO, the videos and their segments are indexed [25, 26] with the extracted metadata.

The paper complements our earlier papers about WARMEMOIRSAMPO: [27] presents the general concept of publishing, searching, and watching the interview videos using Linked Data while [19] gives an overview of the data and text processing and of building the semantic portal to make use of the data.

The paper is structured as follows. Section 2 presents related work relating to the topics of this research. Section 3 discusses the data used in WARMEMOIRSAMPO. Section 4 explains the implementation of NER and NEL for building the facet ontologies, which are evaluated in Section 5. In Section 6, the ontologies are shown in use in the faceted search of the WARMEMOIRSAMPO portal. Section 7 summarizes the results of the paper.

¹<https://www.getty.edu/research/tools/vocabularies/>

²The portal: <https://sotamuistot.arkisto.fi/>

2. Related Work

Digitized cultural heritage documents and artefact collections have inspired many scholars in creating applications for searching, browsing, and recommending data. Typically the dataset metadata is utilized in the search applications to find artefacts based on their metadata similarly [28, 29, 30]. In addition, enriching the metadata using knowledge extraction methods is sometimes possible depending on the collection. For instance, Europeana³ collections consist of more than 50 million CH objects provided by partner institutions across Europe [31, 32]. The Europeana portal provides users with tools to find interesting objects from their digitized collections. This includes also utilizing named entities to recommend content. Similarly, NER has been used in archaeology and war history to build data search applications [33, 34]. In the case of archeology, in addition to using named entities, Brandsen et al. [33] have evaluated BERT-based NER for information retrieval with an archaeological text collection, including advanced search capabilities.

There are several NER tools and models for extracting named entities from Finnish texts. Lately, the three most notable tools are StanfordNER [35], FiNER [36], and FinBERT’s NER models [37]. Out of the three, the BERT-based models for Finnish are evaluated as the most accurate so far [37, 38, 39].

On lemmatization, the more traditional method is based on finite state transducers (FST), for which Omorfi (Open morphology for Finnish) [40] is a prime example for the Finnish language; another one, Voikko⁴, has been in development for many years. A newer method, which uses deep neural networks, is exemplified by the TurkuNLP Neural Parser pipeline [41]. WARMEMOIRSAMPO employs both methods, using FST to correct errors made by neural networks.

3. WarMemoirSampo Data

The WARMEMOIRSAMPO knowledge graph is built from source data that was available in spreadsheet format (CSV). The source data consists of video and interview metadata with textual descriptions of video contents. The core component of the metadata are spreadsheet tables with timestamped textual descriptions of the things being discussed on the interview videos, divided into temporal segments of various lengths. The free-form notes in Finnish and the video segmentation are created originally at the time of the interviews, and later transformed into structured data. The videos and their metadata were provided by the veteran organization Tammenlehvän Perinneyhdistys and the National Archives of Finland.⁵

The interviews were carried out in Finnish, which is more challenging for natural language processing than for example English: its rich morphology⁶ is burdensome for lemmatization and Named Entity Recognition (NER). Currently, Finnish is one of the most well-resourced

³<https://www.europeana.eu/>

⁴<https://voikko.puimula.org/>

⁵More information about the WARMEMOIRSAMPO project can be found at: <https://seco.cs.aalto.fi/projects/war-memoirs/en>

⁶Finnish contains around 15 cases and various suffixes, which result in a number of surface forms that may reach well over 2000 for a single word.

languages for Natural Language Processing (NLP) [42]. However, since the interviews were realized in colloquial, dialectal Finnish, not even the state-of-the-art speech-to-text algorithms proved robust enough to cope with them: according to our test, only the standardized *general language* (*yleiskieli*) produces adequate results presently, in contrast to colloquial and regional variants⁷.

The free-form notes made by the interviewers during the conversations were used as the primary textual data. They do vary in scope and were not originally meant to be used as sources of information the way WARMEMOIRSAMPO requires. Their sentences are not necessarily complete and grammatically correct; abbreviations also abound. This aggravates the challenge posed by Finnish, especially regarding lemmatization and entity recognition.

The source data also contains interviewees' name, date and place of birth, length and place of the interviews, links for the respective interview videos on the YouTube platform, and other metadata. The notes for each interview are divided into rows, roughly corresponding to a sentence; they also feature a timestamp related to their location in the YouTube video. These timestamps are nevertheless not unique, since several consecutive rows usually present the same timestamp, marking the start of the first row in the group. For WARMEMOIRSAMPO these rows were grouped together, resulting in several minutes long video segments, which became our main data unit for this project.

4. Building the Facet Ontologies

The facet ontologies are mostly built by applying NER to the source data, which extracts categorized entities mentioned in the interview notes. This approach is further advanced by applying two separate pre-existing NEL tools on the interview notes, which link to Wikidata and to the WarSampo LOD service on Finnish Second World War history [29, 23] based on the mentions of known entities. The idea for using the NEL approaches in addition to NER, is to be able to show links to further information about the topics that are being discussed in the interviews at a certain time, and also to enrich entity information with alternative labels to be able to automatically reconcile known duplicate entities. For example, a person can be referred to with multiple names which is impossible for NER to handle, as is the case for Mr. "Aarne Juutilainen", also known and mentioned as "Saharan kauhu" ("The dread of Sahara"). After the NER and NEL steps, the created entities are reconciled into six separate lightweight ontologies. This process is depicted in Fig. 1.

4.1. NER Entities and Categories

Named entity recognition was performed using the Secompling library⁸, which aims at providing an integrated interface for various Natural Language Processing (NLP) tools in Finnish. The module responsible for lemmatization and NER employs two tools developed by the TurkuNLP research group⁹: Neural Parser pipeline [43] and Finnish NER [37], which are based on the BERT

⁷There is an ongoing effort to collect different kinds of Finnish speech: www.lahjoitapuhetta.fi

⁸<https://version.aalto.fi/gitlab/seco/secompling>

⁹<https://turkunlp.org/>

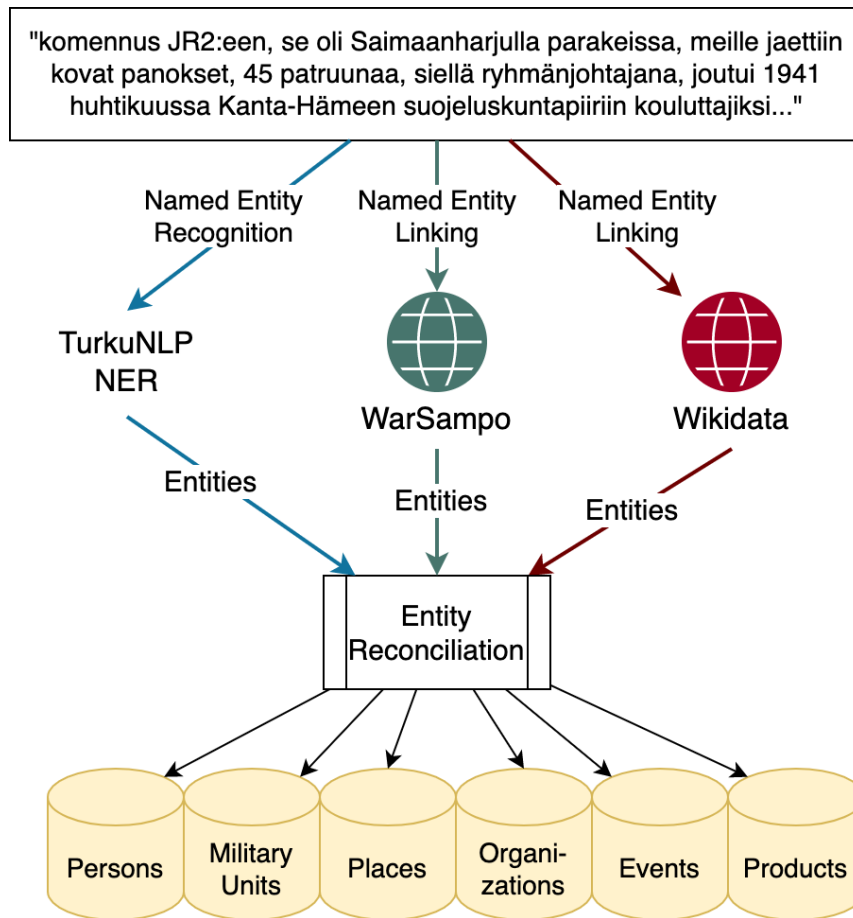


Figure 1: The NER approach is complemented by two separate NEL approaches, using pre-existing NEL solutions, which link to WarSampo and Wikidata. After this the entities are reconciled and six separate lightweight ontologies are created for different domain categories.

model FinBERT [38], also developed by the TurkuNLP group. The parser outputs CoNLL-U¹⁰ documents, which include lemma, part-of-speech (POS) tags, and dependency relations. The NER tool labels each word with an *IOB* tag to mark the *Inside*, *Outside* or *Beginning* of the entities, and a category tag.

Seempling combines the results obtained via both tools in order to fix errors made by either one. For example, at times their tokenization differs: a word that is split by the NER tool is not split by the parser. This may create conflicts if the latter marks it as a proper name but the former does not provide tags for any of its parts. On the other hand, a single entity may be assigned different tags. Wrongly tokenized words may also affect the quality of the POS tags. It is possible to fix some of these errors by aligning the results and combining them heuristically. A total of 1220 documents were altered to produce better parsing. In almost all cases, tokens

¹⁰<https://universaldependencies.org/format.html>

were split so that they started with a punctuation mark and were assigned the wrong POS tag. Separating these punctuation marks improved the results.

Aligning the results also provides lemmatized forms for the recognized named entities. This is indispensable, since the NER tool only tags words found in the text without analyzing them any further. However, the morphological richness of Finnish means that words very often do not appear in their basic form, so they must be lemmatized. The lemmatization module in Secompling uses the third-party libraries Voikko and uralicNLP (whose backend is Omorfi) in order to check and possibly fix the basic forms provided by the parser, using their POS as input.

Out of the 18 categories recognized by the NER tool, only the following were used as classes: Person, Product, Organization, Event, and a combination of Location (LOC), Facilities (FAC) and Geopolitical entities (GPE) as Place. An attempt is made to internally link person entities containing only one word (presumably either first name or surname) to corresponding full-named entities previously detected in the same text; failing that, they are ignored. Moreover, military units were systematically removed from the results, since the recognition provided by the Warsa-linkers tool used in WarSampo was deemed sufficient for this project.

Specific lemmatization heuristics are applied to named entities, since they are often formed by more than one word with different inflections¹¹. A simple ad-hoc evaluation with 100 random entities indicates that the Neural parser has a lemmatization accuracy of 0.84 when it comes to named entities, which is a significant depart from the scores of 0.951–0.972 reported in [41] for datasets of texts in Finnish. This is probably due to the overall lack of multi-word entity lemmatization and the typographic errors in the texts. Secompling raises this entity-specific score to 0.96. However, at this point Secompling is not able to perform external linking, therefore it cannot differentiate accurate named entities from errors in recognition and/or lemmatization. As a compensatory measure, manual correction of entities is used, so that entities can be removed, or their lemma and/or category can be changed.

During the transformation a set of correction rules were applied to the results in order to remove and correct the entities to be shown in facets: Ambiguous single-name person entities were discarded, which amount to over 500. A total of 336 entities were blacklisted, including 182 organizations (which overlap significantly with locations, when the entities are facilities such as hospitals, airports, etc.; they might be recognized as either one), 63 locations (mostly related to military units), 30 events (many overlap with locations) and 61 products (mostly deemed irrelevant). For four entities the category was changed, in 221 cases labels were changed, and in 45 cases both the category and the label had to be changed.

4.2. Parallel NEL Linking to Wikidata and WarSampo

In addition to the NER process, two separate NEL processes are used directly on the source data. The NEL processes are based solely on the texts in the interview notes.

WarSampo. The WarSampo system includes a data and ontology infrastructure for representing Finland in World War II. The WarSampo knowledge graph (KG) [23] consists of around 100 000 persons, 51 000 places, 16 000 military units, 166 000 photographs, 26 000 war diaries, and 3000 war veteran memoir articles, among others, with large amounts of links between

¹¹For example, genitive followed by nominative, as in *Helsingin päärautatieasema* 'Helsinki Central Station'

entities. The persons contains the casualty register [44] of the National Archives of Finland, containing detailed information about all 94 700 people killed in action in Finland during WW2, as well as a person register of all of the 4200 Finnish prisoners of war [45] and 5600 notable individuals from other data sources (Wikipedia, etc.) [46]. WarSampo also uses the official Place Name Register of Finland by National Survey as an external domain ontology, served on a separate SPARQL endpoint.

In WarSampo, textual event and photograph descriptions were linked to persons, places, and military units using the *Warsa-linker* tool [47]. In WARMEMOIRSAMPO, Warsa-linker was re-used to link mentions of persons, places, and military units to corresponding entities in the WarSampo KG. After the NEL phase, metadata of the linked entities was pulled with SPARQL queries from the WarSampo KG, and new entities were created to WARMEMOIRSAMPO based on the retrieved metadata, with links to the original entities.

Wikidata. In order to link and identify entities with the Wikidata knowledge base, the named entity linking tool *Nelli* [48, 49] was utilized. In the WARMEMOIRSAMPO context, Nelli is configured to first lemmatize texts using the Turku Neural Parser pipeline and then to link the results with ARPA [50]. ARPA is a configurable entity linking tool that can be configured for different services; in the case of WARMEMOIRSAMPO, the tool was configured to link places, units, and people to Wikidata.

In the interviews, the veterans mention some of their comrades in arms and superiors. It is easy to identify and link the most notable officers, however, disambiguating regular soldiers is hard. The latter are often mentioned by only first name and/or surname and there can be several namesakes, so that more information would have been necessary to differentiate them. Therefore, the linking strategy is centered around notable figures present in Wikidata.

In terms of place linkage, several Wikidata ARPA configurations are used to link places such as continents, countries, cities, as well as smaller places such as towns and villages. It can be challenging to match former Finnish place names to Wikidata due to varying practices in classifying place entities (e.g., towns or urban settlements) or missing labels, e.g., entities *Paanajärvi* or *Muolaa* in Wikidata. Moreover, places that were annexed to Russia often changed their names from Finnish to Russian, and some place names were lacking their original names mentioned in the interviews. When such clear shortcomings in Wikidata were identified by manual inspection, we contributed to Wikidata by adding the missing Finnish names.

4.3. Reconciling Entities Between NER and NEL

After applying NER and two different NEL tools, there are a large number of duplicate entities. Most of these are due to the different processes finding entities for the same words in the text, whereas in some cases one entity is referred to with multiple names. The numbers of entities found by class and source are given in Table 1. The “Entity Class” column corresponds to the entity class, “NER” corresponds to the number of entities created with the NER process, “WarSampo” and “Wikidata” correspond to the number of entities created from the respective NEL sources.

The generated named entities from different steps are disambiguated between the different data sources based on 1) matching the sets of URIs received for them from LOD data sources, and 2) matching entity types and names. The entities are then reconciled by merging the found

Table 1

Numbers of entities from the NER and NEL processes separated by entity class and including the final number of reconciled entities from the separate processes.

Entity Class	NER	WarSampo	Wikidata	Reconciled Entities
Place	1613	961	350	1840
Person	209	159	42	310
Organization	594	0	16	610
Military Unit	0	112	5	117
Event	66	0	0	66
Product	51	0	0	51

duplicates. The numbers of disambiguated and reconciled named entities per entity class are shown in Table 1 column “Reconciled Entities”.

5. Evaluating the Facet Ontologies

The six facet ontologies consist of entities of the corresponding entity class with label and source information, but no links between the entities. The entities from the NEL processes also contain links to the entities in external systems and metadata pulled from them. The facet ontologies by entity class are:

1. **Places** contain 1840 place entities (found in the notes of all 159 interviews) from NER, WarSampo, and Wikidata. NER recognizes a broad range of places into this category, including countries, municipalities, cities, towns, geographical formations like islands and rivers, and buildings such as churches, train stations, and airfields. The WarSampo linking links to all place types in the WarSampo KG, consisting of municipalities, cities, and various types of smaller geographical locations. The Wikidata linking links to continents, countries, cities, and municipalities. The place entities with most links from the interviews are important locations in the war: The Karelian Isthmus, Finland, Helsinki, Vyborg, Svir. On the other hand, there is a long tail of places with varying relevance and some clear errors.
2. **Persons** contain 310 person entities (found in the notes of 126 out of 159 interviews) from NER, WarSampo, and Wikidata. The most referenced entities are prominent figures during the wartime, including, e.g., the Finnish commander-in-chief C.G.E. Mannerheim, Joseph Stalin, and Finnish commissioned officers. Due to the rather small size, it was possible to clean this facet ontology of erroneous entities. However, one of the most referenced entities “Ehnrooth” is actually a group of several prominent Finnish commissioned officers with that surname. In some cases the first names are used as well and these are disambiguated to a specific Ehnrooth. In other cases, such as mentions containing only a surname, it is not possible to disambiguate person entities, except in cases like “Mannerheim”, which can be disambiguated with high confidence.
3. **Military Units** contain 117 entities (found in the notes of 122/159 interviews) from WarSampo and Wikidata. The entities are mostly different infantry regiments, but also contain some often referenced general entities like reserve officer school, headquarters,

and air force. As they come only from applying NEL to curated contents, there are no clear errors present and the entities don't seem to contain duplicates.

4. **Organizations** contain 610 entities (found in the notes of 155/159 interviews) from NER and Wikidata. Almost all of the entities come from NER and the scope of the organizations seems to be very broad, including a few often referenced entities like the White Guard, Waffen-SS, and the Finnish War Veteran Association. In addition to the few entities with a large amount of mentions, there is a long tail of organizations, such as schools, companies, museums, travel agencies, and clubs, for many of which it is difficult to ascertain their identity without listening to how they are referenced in the actual interviews and trying to find information from external sources. Some of these seem like errors in NER categorization, but could be names of, e.g., companies instead. Due to the high number of entities and the effort it takes to study their identities, this ontology was left rather unfinished, but with some of the clear errors removed. Also the NER assigned many military units into this class, which were removed due to them being found best with the WarSampo linking.
5. **Events** contain 66 entities (found in the notes of 151/159 interviews) from NER. The most referenced entities are very relevant: Winter War, Continuation War, Lapland War, and Civil War. Half of the entities are battles, usually named after the place of occurrence. As the number of entities is low, it was easy to manually curate the list.
6. **Products** contain 51 entities (found in the notes of 25/159 interviews) from NER. These have a low number of references overall, and there are no entities with high number of references. The ones with most references (3) are military aircraft makes and models. There are also car brands, medals, and a variety of entities, which, like in the case of organizations, would require plenty of effort to study their identities, and some of them are probably errors.

Due to the fully separate NER and NEL processes, in some cases we ended up creating multiple named entities from one entity mention in the text. This occurs when locations like "Iisalmi church" are mentioned, of which we get named entities for both "Iisalmi" (a municipality) and "Iisalmi church".

After the ontologies were initially created and the WARMEMOIRSAMPO PORTAL set up, it was possible to observe the facet ontologies in use and to spot errors and irrelevant entities. Several iterations of adjusting the entity blacklists were done at this point. In this iterative process of improving the ontologies, it was found out that in some cases the inspection of whether an entity is having the right class required listening to the interviews to get the context what is being discussed. This would have been different had we been working with the actual transcriptions of the interviews instead of concise and varying notes.

The results of this process were evaluated by inspecting the final named entities assigned to twenty random interview segments. They contain a total of 99 recognized entities. Out of these, 88 were exact identifications and three entities were doubly recognized, as explained above. Additionally, two entities were mistakenly linked¹², two mistakenly recognized and four received the wrong category; and five were not recognized. This adds up to a Precision of 0.919

¹²One entity was disregarded, since a typographical error led to a wrong identification: *Turkin saari* instead of *Turkinsaari*. Entity without links (a total of 5) were considered correct.

(or 0.889 when disregarding double recognitions) and a Recall of 0.875 (0.846), and an F1 score of 0.897 (0.867).

A more robust entity linking system could avoid mistaken links, but due to disambiguation effects it could also help avert erroneous entities and categories. Improving the lemmatization of named entities could also raise this score in case some entities are not recognized due to their surface form.

6. Faceted Search User Interface

Sampo-UI [12] is a JavaScript framework for building web-based user interfaces for KGs published in SPARQL endpoints. The ideas and design choices behind Sampo-UI have been gradually developed in the context of the general Sampo model [51], which have guided the implementation of a series of public semantic portal prototypes¹³ since 2002, mostly in the cultural heritage domain. As faceted search is the key search paradigm in Sampo-UI, it was a natural tool of choice for implementing a user interface that utilizes the new facet ontologies which are published within the WARMEMOIRSAMPO KG.

In the current version of Sampo-UI, a new user interface is implemented by writing a set of JavaScript Object Notation (JSON) configuration files¹⁴ accompanied with corresponding SPARQL queries¹⁵, which are then automatically used as input parameters for assembling a fully functional modern web application called WARMEMOIRSAMPO PORTAL¹⁶, which uses data directly from the WARMEMOIRSAMPO SPARQL endpoint¹⁷.

The simplest solution for delivering the interview videos to users would be a clickable list of videos with thumbnails, possibly with an option to search for the titles of the videos. The aim of the WARMEMOIRSAMPO PORTAL is to go far beyond this by enabling search functionalities that can dig out specific segments inside the long interviews. This way the user can start watching an interview directly at a point where an entity of interest (e.g. person, place, military unit) is mentioned. In addition to enhanced search tools, the underlying facet ontologies make it possible to provide rightly timed contextual links when the user is watching the video. Using the links the user can learn more¹⁸ about the entities mentioned in the current segment of the interview video.

At the front page of the portal the user can choose between three distinct faceted search perspectives:

1. *Interviews* perspective for searching whole interviews using a free combination of 10 facets.
2. *Segments* perspective for searching specific segments inside the interviews with same facets as in the Interviews perspective.

¹³See the full list of semantic portals based on the Sampo model at <https://seco.cs.aalto.fi/applications/sampo>.

¹⁴<https://github.com/SemanticComputing/veterans-web-app/tree/master/src/configs>

¹⁵https://github.com/SemanticComputing/veterans-web-app/tree/master/src/server/sparql/veterans/sparql_queries

¹⁶The portal is published at <https://sotamuistot.arkisto.fi>

¹⁷The SPARQL endpoint is available at <https://ldf.fi/warmemoirsampo/sparql>

¹⁸Links to WarSampo and Wikipedia are provided as explained in subsection 4.2.

3. *Directory* perspective for searching all automatically recognized named entities mentioned on in the interview notes, and accessing the segments where the entity is mentioned.

Fig. 2 illustrates how the five video segments in which the place *Äänislinna (Petrozavodsk)* and the military unit *JR 50* are both mentioned can be found using two facet selections. In addition to faceted search, the underlying facet ontologies can be used for implementing exploratory search functionalities. As an example of this, Fig. 3 portrays how the 4566 distinct geographical references within the segments are plotted on an interactive map. Clicking a marker on the map opens a popup which provides links to all segments where the place is mentioned.

Faceted search offers a few key benefits to the users of WARMEMOIRSAMPO that would be missing if only traditional text search would be available. Notably, with faceted search the user does not need to know what she is searching for. This makes it possible to quickly find unexpected interesting things in the interviews. For example while browsing through the values of the *Mentioned Product* facet the user may pay attention that the famous Finnish spreadable cheese *Koskenlaskija* is mentioned in one interview segment. By clicking the search result the user can watch the segment, which happens to include a story about *Koskenlaskija* cheese saving a person's life¹⁹.

Because the faceted search shows the user the hit counts for each option in the facets, the user can also get a general idea about the things that the interviewees are talking about with almost a single glance. For example, in the *Mentioned place* facet most hits are, unsurprisingly,

¹⁹The interviewed veteran gave his comrade a *Koskenlaskija* cheese packet and the comrade moved away from his normal battle station to eat the cheese, just before a bomb hit and killed the other person in that station.

The screenshot shows the WARMEMOIRSAMPO portal interface. At the top, there's a navigation bar with 'Haastattelut', 'Haastattelujen kohdat', 'Hakemisto', 'Palautte', 'Info', and 'Ohjeet'. The main header reads 'Haastattelujen kohdat' and 'Results: 5 segments that mention the place Äänislinna AND military unit JR 50'. Below this, there's a search bar and a table of results. On the left, there are several facet panels: 'Aktiiviset suodattimet' (Active filters), 'Paikka (maininta)' (Location), 'Henkilö (maininta)' (Person), and 'Joukko-osasto (maininta)' (Unit). A red arrow points to the 'Paikka (maininta)' facet with the text '10 facets for semantic searching'. The table of results shows five segments, each with a video thumbnail, a title, a description, and a list of associated facets and their hit counts.

Figure 2: Faceted search for segments inside interview videos in the WARMEMOIRSAMPO portal. The left-hand side shows the facets with hit counts and the search results are displayed on the table view in the middle.

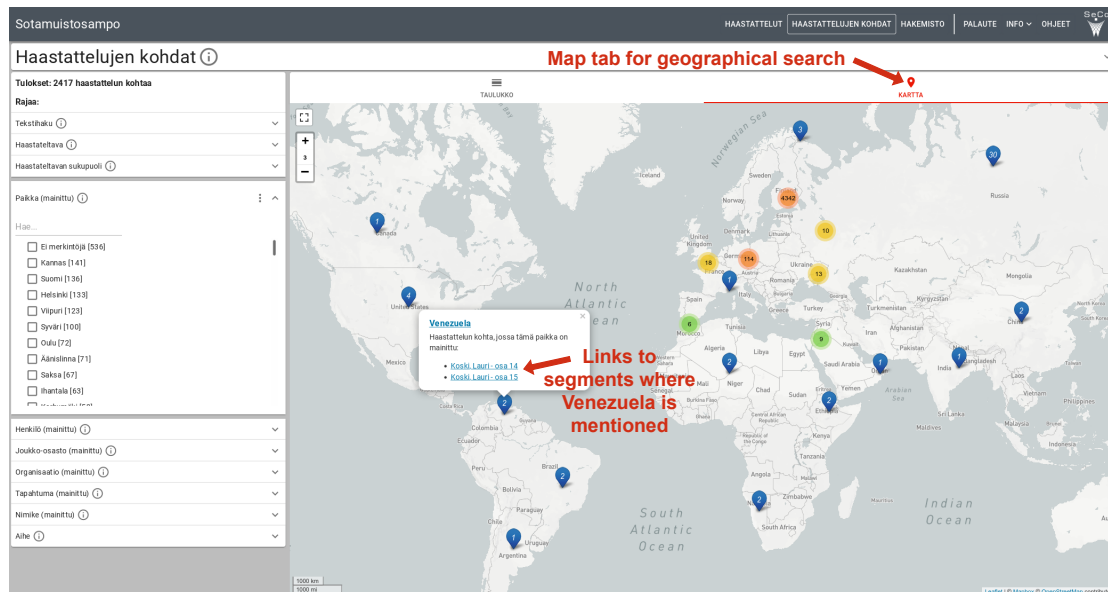


Figure 3: An exploratory search scenario in the WARMEMOIRSAMPO portal where the map view is selected to access the video segments that mention a specific place.

certain places in Karelia where most important events of the war took place, and where most of the soldiers were fighting. Germany is also mentioned quite often, interestingly more often than Russia or Soviet Union. On the other hand, when looking at places with only one or few hits, the user can find some surprising and interesting entities, such as Brazil mentioned in an interview. While this kind of comparison can be useful, it should not be taken too seriously, as the entities are disambiguated only based on the entity name and class. For example, for some place names, there are actually multiple places with the same name and in these cases they are grouped together as one entity. The main goal of finding named entities in WARMEMOIRSAMPO is to facilitate the search, which makes these issues less relevant than they would be in some other cases.

7. Conclusion

In this paper, we have presented the idea of building lightweight ontologies for faceted search with NER, and shown how state-of-the-art NER and NEL tools were used in the WARMEMOIRSAMPO system to create good enough lightweight ontologies to provide facet options in a faceted search user interface. This approach follows the trend in Semantic Web research to shift from the heavy use of formal semantics towards leveraging the collection of distributed, heterogeneous data using light-weight ontologies [52].

In the entity recognition task, we achieved an F1 score of 0.90 for a small sample of 20 interview segments, discounting wrong links, entities and categories. Our lemmatization task had accuracy of 0.96 within a sample of 100 entities. Improving lemmatization could have

positive knock-on effects on NER; so could enhancing the linking system.

We have shown how to find entities from Finnish textual data and categorize them with high enough recall and precision to be useful for building facet ontologies in practice, without involving considerable manual domain ontology engineering. However, the created facet ontologies would still benefit from some post-processing to remove erroneous and non-relevant entities, and from organizing them into hierarchies, which would facilitate search using higher level categories. However, in our case this was deemed not necessary as the number of entities is not very large.

The WARMEMOIRSAMPO data is relatively small, containing ca. 323 000 triples. The tools used in this study would scale up for a few times larger datasets without any effort. However, the NER tool, which is the key component for our approach, would scale even better. The amount of manual work required with the facet ontologies might also increase as the source data size increases, but the extent of this would depend on the data. In the future, we will further research this approach of using NER to extract entities for building facet ontologies in different contexts.

In the cultural heritage domain, transforming data into Linked Data often requires manually curated data transformation processes that align the data with the used data models and ontologies according to defined rules [6, 8, 53, 23]. The ontologies are either manually engineered or in some cases pre-existing vocabularies and ontologies may be used that fit the needs of the data publication. The approach of building facet ontologies with NER presented in this paper could significantly reduce the effort required to create Linked Data from textual datasets. As facet ontologies can be built from data bottom-up, it becomes easier to implement faceted search for searching, browsing and visualizing the data.

Acknowledgments

Ilpo Murtovaara, Markus Merenmies, and Kare Salonvaara contributed in creating the videos and their metadata used in the work of this paper. Tammenlehvän Perinneyhdistys ry with the National Archives of Finland funded our work. The work is partly related and funded by the EU project InTaVia: In/Tangible European Heritage²⁰, and the EU COST action Nexus Linguarum²¹ on linguistic data science. The authors acknowledge CSC – IT Center for Science, Finland, for providing for computational resources.

References

- [1] M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, K.-P. Yee, Finding the flow in web site search, *Communications of the ACM* 45 (2002) 42–49. doi:10.1145/567498.567525.
- [2] D. Tunkelang, *Faceted Search, Synthesis lectures on information concepts, retrieval, and services*, Morgan & Claypool Publishers, 2009.
- [3] A. S. Pollitt, *The key role of classification and indexing in view-based searching*, Technical Report, Centre for Database Access Research, University of Huddersfield, 1998. URL: <http://www.ifla.org/IV/ifla63/63polst.pdf>.

²⁰<https://intavia.eu/>

²¹<https://nexuslinguarum.eu/the-action>

- [4] G. M. Sacco, Dynamic taxonomies for intelligent information access, in: M. Khosrow-Pour, D.B.A. (Ed.), *Encyclopedia of Information Science and Technology*, Third Edition, Hershey, PA: IGI Global, 2015, pp. 3883–3892. doi:10.4018/978-1-4666-5888-2.ch382.
- [5] E. Hyvönen, S. Saarela, K. Viljanen, Application of ontology techniques to view-based semantic search and browsing, in: *The Semantic Web: Research and Applications. Proceedings of the First European Semantic Web Symposium (ESWS 2004)*, 2004. doi:10.1007/978-3-540-25956-5_7.
- [6] E. Hyvönen, E. Mäkelä, M. Salminen, A. Valo, K. Viljanen, S. Saarela, M. Junnila, S. Kettula, MuseumFinland—Finnish museums on the Semantic Web, *Journal of Web Semantics* 3 (2005) 224–241. doi:10.1016/j.websem.2005.05.008.
- [7] E. Mäkelä, E. Hyvönen, T. Sidoroff, View-based user interfaces for information retrieval on the Semantic Web, in: *Proceedings of the ISWC 2005 Workshop on End User Semantic Web Interaction*, CEUR Workshop Proceedings, 2005. Vol. 172.
- [8] G. Schreiber, A. Amin, L. Aroyo, M. van Assem, V. de Boer, L. Hardman, M. Hildebrand, B. Omelayenko, J. van Osenbruggen, A. Tordai, J. Wielemaker, B. Wielinga, Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator, *Journal of Web Semantics* 6 (2008) 243–249. doi:10.1016/j.websem.2008.08.001.
- [9] Y. Tzitzikas, N. Manolis, P. Papadakos, Faceted exploration of RDF/S datasets: a survey, *Journal of Intelligent Information Systems* 48 (2017) 329–364. doi:10.1007/s10844-016-0413-8.
- [10] M. Hildebrand, J. Van Ossenbruggen, L. Hardman, /facet: A browser for heterogeneous Semantic Web repositories, in: I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, L. M. Aroyo (Eds.), *The Semantic Web – ISWC 2006: 5th International Semantic Web Conference*, volume 4273 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2006, pp. 272–285. doi:10.1007/11926078_20.
- [11] M. Koho, E. Heino, E. Hyvönen, SPARQL Faceter – client-side faceted search based on SPARQL, in: *Joint Proceedings of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop*, CEUR Workshop Proceedings, 2016. URL: <http://www.ceur-ws.org/Vol-1615>, vol. 1615.
- [12] E. Ikkala, E. Hyvönen, H. Rantala, M. Koho, Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces, *Semantic Web – Interoperability, Usability, Applicability* 13 (2022) 69–84. doi:10.3233/SW-210428.
- [13] C. Bizer, T. Heath, T. Berners-Lee, Linked Data – the story so far, *International Journal on Semantic Web and Information Systems (IJSWIS)* 5 (2009) 1–22. doi:10.4018/jswis.2009081901.
- [14] T. R. Gruber, A translation approach to portable ontology specifications, *Knowledge acquisition* 5 (1993) 199–220.
- [15] M. C. Suárez-Figueroa, A. Gómez-Pérez, M. Fernández-López, The NeOn methodology for ontology engineering, in: *Ontology Engineering in a Networked World*, Springer, Berlin, Heidelberg, 2012, pp. 9–34. doi:10.1007/978-3-642-24794-1_2.
- [16] C. Fluit, M. Sabou, F. Van Harmelen, Supporting user tasks through visualisation of light-weight ontologies, in: *Handbook on Ontologies*, Springer, Berlin, Heidelberg, 2004, pp. 415–432.
- [17] F. Giunchiglia, I. Zaihrayeu, Lightweight ontologies, in: L. Liu, M. T. Özsu (Eds.), *Ency-*

- yclopedia of Database Systems, Springer, Boston, MA, 2009, pp. 1613–1619. doi:10.1007/978-0-387-39940-9_1314.
- [18] L. Buitinck, M. Marx, Two-stage named-entity recognition using averaged perceptrons, in: G. Bouma, A. Ittoo, E. Métais, H. Wortmann (Eds.), *Natural Language Processing and Information Systems*, volume 7337 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2012, pp. 171–176.
- [19] R. Leal, H. Rantala, M. Koho, E. Ikkala, M. Merenmies, E. Hyvönen, WarMemoirSampo: A semantic portal for war veteran interview videos, in: 6th Digital Humanities in Nordic and Baltic Countries Conference, short paper, 2022. Forth-coming.
- [20] W. Shen, J. Wang, J. Han, Entity linking with a knowledge base: Issues, techniques, and solutions, *IEEE Transactions on Knowledge and Data Engineering* 27 (2014) 443–460. doi:10.1109/TKDE.2014.2327028.
- [21] B. Hachey, W. Radford, J. Nothman, M. Honnibal, J. R. Curran, Evaluating entity linking with Wikipedia, *Artificial Intelligence* 194 (2013) 130–150. doi:10.1016/j.artint.2012.04.005.
- [22] R. C. Bunescu, M. Pasca, Using encyclopedic knowledge for Named Entity Disambiguation, in: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, volume 6, Association for Computational Linguistics, Trento, Italy, 2006, pp. 9–16.
- [23] M. Koho, E. Ikkala, P. Leskinen, M. Tamper, J. Tuominen, E. Hyvönen, WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data, *Semantic Web – Interoperability, Usability, Applicability* 12 (2021) 265–278. doi:10.3233/SW-200392.
- [24] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, D. Vrandečić, Introducing Wikidata to the Linked Data web, in: *The Semantic Web – ISWC 2014*, volume 8796 of *Lecture Notes in Computer Science*, Springer, Cham, 2014, pp. 50–65. doi:10.1007/978-3-319-11964-9_4.
- [25] J. Hunter, R. Iannella, The application of metadata standards to video indexing, in: C. Nikolaou, C. Stephanidis (Eds.), *Research and Advanced Technology for Digital Libraries*, Springer, Berlin, Heidelberg, 1998, pp. 135–156.
- [26] C. Ribeiro, M. L. Mucheroni, Dynamic indexation in video metadata, *Procedia-Social and Behavioral Sciences* 73 (2013) 551–555.
- [27] E. Hyvönen, E. Ikkala, M. Koho, R. Leal, H. Rantala, M. Tamper, How to search and contextualize scenes inside videos for enriched watching experience: Case stories of the Second World War veterans, in: *Proceedings of the 19th Extended Semantic Web Conference (ESWC 2022)*, Poster and Demo papers, Springer, 2022. URL: <https://seco.cs.aalto.fi/publications/2022/hyvonen-et-al-wms-2022.pdf>, forth-coming.
- [28] S. Dumont, correspSearch – connecting scholarly editions of letters, *Journal of the Text Encoding Initiative* 10 (2016). doi:10.4000/jtei.1742.
- [29] E. Hyvönen, E. Heino, P. Leskinen, E. Ikkala, M. Koho, M. Tamper, J. Tuominen, E. Mäkelä, WarSampo data service and semantic portal for publishing Linked Open Data about the Second World War history, in: *The Semantic Web – Latest Advances and New Domains (ESWC 2016)*, Springer, 2016, pp. 758–773. doi:10.1007/978-3-319-34129-3_46.
- [30] E. Hyvönen, E. Mäkelä, T. Kauppinen, O. Alm, J. Kurki, T. Ruotsalo, K. Seppälä, J. Takala, K. Puputti, H. Kuittinen, K. Viljanen, J. Tuominen, T. Palonen, M. Frosterus, R. Sinkkilä,

- P. Paakkari, J. Laitio, K. Nyberg, CultureSampo – a national publication system of cultural heritage on the Semantic Web 2.0, in: Proceedings of the 6th European Semantic Web Conference (ESWC 2009), Springer, 2009.
- [31] S. Gordea, M. L. Paramita, A. Isaac, Named entity recommendations to enhance multilingual retrieval in Europeana.eu, in: Foundations of Intelligent Systems. ISMIS 2020., volume 12117 of *Lecture Notes in Computer Science*, Springer, Cham, 2020, pp. 102–112. doi:10.1007/978-3-030-59491-6_10.
- [32] V. Petras, T. Hill, J. Stiller, M. Gäde, Europeana - a search engine for digitised cultural heritage material, *Datenbank-Spektrum* 17 (2017) 41–46.
- [33] A. Brandsen, S. Verberne, K. Lambers, M. Wansleben, Can BERT dig it? – named entity recognition for information retrieval in the archaeology domain, *Journal on Computing and Cultural Heritage* (2021). doi:10.1145/3497842.
- [34] M. Tamper, P. Leskinen, E. Ikkala, A. Oksanen, E. Mäkelä, E. Heino, J. Tuominen, M. Koho, E. Hyvönen, AATOS – a configurable tool for automatic annotation, in: Proceedings, Language, Data and Knowledge (LDK 2017), volume 10318 of *Lecture Notes in Computer Science*, Springer, Cham, 2017, pp. 276–289. doi:10.1007/978-3-319-59888-8_24.
- [35] J. R. Finkel, T. Grenager, C. D. Manning, Incorporating non-local information into information extraction systems by Gibbs sampling, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05), June, 25–30, 2005, University of Michigan, Ann Arbor, Michigan, USA, Association for Computational Linguistics, 2005, pp. 363–370. doi:10.3115/1219840.1219885.
- [36] T. Ruokolainen, P. Kauppinen, M. Silfverberg, K. Lindén, A Finnish news corpus for Named Entity Recognition, *Language Resources and Evaluation* 54 (2020) 247–272.
- [37] J. Luoma, M. Oinonen, M. Pyykönen, V. Laippala, S. Pyysalo, A broad-coverage corpus for Finnish Named Entity Recognition, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 4615–4624. URL: <https://aclanthology.org/2020.lrec-1.567>.
- [38] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, S. Pyysalo, Multilingual is not enough: BERT for Finnish, 2019. URL: <https://arxiv.org/abs/1912.07076>. arXiv:1912.07076, arXiv.
- [39] T. Ruokolainen, K. Kettunen, À la recherche du nom perdu—searching for named entities with Stanford NER in a Finnish historical newspaper and journal collection, in: 13th IAPR International Workshop on Document Analysis Systems, 2018.
- [40] T. A. Pirinen, Development and use of computational morphology of Finnish in the open source and open science era: Notes on experiences with Omorfi development., *SKY Journal of Linguistics* 28 (2015) 381–393.
- [41] J. Kanerva, F. Ginter, T. Salakoski, Universal lemmatizer: A sequence-to-sequence model for lemmatizing universal dependencies treebanks, *Natural Language Engineering* 27 (2021) 545–574. doi:10.1017/S1351324920000224.
- [42] M. Hämäläinen, K. Alnajjar, The current state of Finnish NLP, in: Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages, The Association for Computational Linguistics, 2021, pp. 54–61.
- [43] J. Kanerva, F. Ginter, N. Miekka, A. Leino, T. Salakoski, Turku Neural Parser pipeline: An end-to-end system for the CoNLL 2018 shared task, in: Proceedings of the CoNLL 2018

Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, 2018.

- [44] M. Koho, E. Hyvönen, E. Heino, J. Tuominen, P. Leskinen, E. Mäkelä, Linked death — representing, publishing, and using Second World War death records as Linked Open Data, in: E. Blomqvist, K. Hose, H. Paulheim, A. Lawrynowicz, F. Ciravegna, O. Hartig (Eds.), *The Semantic Web: ESWC 2017 Satellite Events*, volume 10577 of *Lecture Notes in Computer Science*, Springer, Cham, 2017, pp. 369–383. doi:10.1007/978-3-319-70407-4_45.
- [45] M. Koho, E. Ikkala, E. Hyvönen, Reassembling the lives of Finnish prisoners of the Second World War on the Semantic Web, in: *Proceedings of the Third Conference on Biographical Data in the Digital Age (BD 2019)*, CEUR Workshop Proceedings, 2020. In press.
- [46] P. Leskinen, M. Koho, E. Heino, M. Tamper, E. Ikkala, J. Tuominen, E. Mäkelä, E. Hyvönen, Modeling and using an actor ontology of Second World War military units and personnel, in: C. d’Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, J. Heflin (Eds.), *The Semantic Web – ISWC 2017: 16th International Semantic Web Conference*, volume 10588 of *Lecture Notes in Computer Science*, Springer, Cham, 2017, pp. 280–296. doi:10.1007/978-3-319-68204-4_27.
- [47] E. Heino, M. Tamper, E. Mäkelä, P. Leskinen, E. Ikkala, J. Tuominen, M. Koho, E. Hyvönen, Named Entity Linking in a complex domain: Case second world war history, in: J. Gracia, F. Bond, J. P. McCrae, P. Buitelaar, C. Chiarcos, S. Hellmann (Eds.), *Language, Data, and Knowledge: First International Conference, LDK 2017*, volume 10318 of *Lecture Notes in Computer Science*, Springer, Cham, 2017. doi:10.1007/978-3-319-59888-8_10.
- [48] M. Tamper, A. Oksanen, J. Tuominen, A. Hietanen, E. Hyvönen, Automatic annotation service appi: Named entity linking in legal domain, in: *The Semantic Web: ESWC 2020 Satellite Events*, Springer, 2020, pp. 110–114. doi:10.1007/978-3-030-62327-2_36.
- [49] M. Tamper, E. Hyvönen, P. Leskinen, Visualizing and analyzing networks of named entities in biographical dictionaries for digital humanities research, in: *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*, Springer, 2019. URL: <https://seco.cs.aalto.fi/publications/2021/tamper-et-al-cicling-2021.pdf>, forth-coming.
- [50] E. Mäkelä, Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text, in: *The Semantic Web: ESWC 2014 Satellite Events*, volume 8798 of *Lecture Notes in Computer Science*, Springer, Cham, 2014, pp. 424–428. doi:10.1007/978-3-319-11955-7_60.
- [51] E. Hyvönen, Digital humanities on the Semantic Web: Sampo model and portal series, *Semantic Web – Interoperability, Usability, Applicability* (2022). URL: <http://www.semantic-web-journal.net/content/digital-humanities-semantic-web-sampo-model-and-portal-series-0>, forth-coming.
- [52] A. Bernstein, J. Hendler, N. Noy, A new look at the Semantic Web, *Communications of the ACM*, New York, NY, USA 59 (2016) 35–37. doi:10.1145/2890489.
- [53] M. Koho, T. Burrows, E. Hyvönen, E. Ikkala, K. Page, L. Ransom, J. Tuominen, D. Emery, M. Fraas, B. Heller, D. Lewis, A. Morrison, G. Porte, E. Thomson, A. Velios, H. Wijsman, Harmonizing and publishing heterogeneous pre-modern manuscript metadata as Linked Open Data, *Journal of the Association for Information Science and Technology (JASIST)* 73 (2021) 240–257. doi:10.1002/asi.24499.