Representing and searching associations in cultural heritage knowledge graphs using faceted search

Heikki RANTALA [a,1], Petri LESKINEN [a], Lilli PEURA [a], and Eero HYVÖNEN [a,b],

[a] *Semantic Computing Research Group (SeCo), Department of Computer Science, Aalto University, Finland*

[b] *University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG), Finland*

ORCiD ID: Heikki Rantala https://orcid.org/0000-0002-4716-6564, Petri Leskinen https://orcid.org/0000-0003-2327-6942, Eero Hyvönen https://orcid.org/0000-0003-1695-5840

Abstract. This paper presents how relations or associations between entities, such as persons and places in cultural heritage knowledge graphs, can be searched and analyzed using faceted search and visualizations. Faceted search using well-formed ontologies allows search and comparison of relative numbers in associations of groups of entities, such as artists from different countries, and can reveal patterns of interest in the data. This papers presents examples of how this can be done in practice, how the associations can be conceptualized in different ways that affect the performance of the search, and how the associations can be analyzed. The concept of faceted relational search is examined through case studies including searching relations in collections of biographies from various European countries, relations in the Union List of Artist Names (ULAN) thesaurus, and relations formed by links between Wikipedia pages of persons.

Keywords. Linked data, Digital Humanities, Art History, Biographies, Relational search, Association search,

1. Introduction

Relational search [1], also referred as semantic association search [2], is a search paradigm where the goal is to find relations or connections between entities usually in a context of an RDF[2] knowledge graph (KG). Typically this means searching how two specified entities, such as persons or places, are related in a KG, or finding what entities are somehow related to one specified entity. This paper presents concept of *faceted relational search* where the focus is on searching the relations themselves and comparing their relative numbers, instead of only finding relations between two individual entities. In this concept the unit searched is still an individual association between entities, but the properties of the entities and the association itself can be used to compare groups of en-

[1] Corresponding Author: Heikki Rantala, heikki.rantala@aalto.fi

[2] `https://www.w3.org/RDF/`

tities, through simply faceted search[3] or visualizations. For example, data can include associations between important persons from different countries, and information about the persons such as their occupation and country of origin. Using occupation and country of origin of the two persons forming the association as facets it can be easily seen if all people from certain country have more connections to a certain other country, or if perhaps people of certain occupation have more connections to one country, while people from other occupation would have more connections to some other country. This is similar to searching and visualizing properties of persons for prosopographical analysis, but the patterns revealed by searching associations can be different than when searching persons. The potential research problems that can be solved with this approach include finding out which artists had the most connections of certain type to other artists, or do people from a country A have more connections to country B than to country C.

The approach of this paper is mainly based the knowledge based approach to relational search originally presented in [4]. There relations between persons and places were searched using a KG based on based on the Finnish National Biobraphy [5,6] created using SPARQL CONSTRUCT queries. This paper presents case studies where the approach is extended to searching relations between persons and applied to the Getty ULAN knowledge graph of artists[3] and InTaVia[4] knowledge graph of biographies from multiple European countries. Third case study uses somewhat different approach and is based on extracting links from Wikipedia combined with information from Wikidata[5] KG.

Web application with faceted search functionality and visualizations were created for each case study. The applications were built based on the open code Sampo-UI [7,8] framework where faceted search is implemented using SPARQL queries. In the approach used by Sampo-UI a search perspective limits the search to instances of one or multiple classes, and the search can be further qualified by using facets. Facets are defined so that they represent a property path from instance of the class of the search perspective to a facet value. Hit counts for facets are calculated and updated with SPARQL query after each new selection is made.

2. Related Work

Generally in relational search the *query* consists of two or more resources, and the task is to find interesting semantic associations between them. The approaches [2] differ, at least, in terms of the query formulation, underlying KG, methods for finding connections, and representation of the results. The concept of relational search has been applied in multiple different fields. In [1] the idea of searching relations is applied to the national security domain and finding potential terrorists. In [9] the concept is applied to medical research and genetics. Relational search has also been applied in the Cultural Heritage field in [10,4]. CultureSampo[6] [11,12] contains an application where connections between two artists can be searched based on data from Getty ULAN KG.

[3] https://www.getty.edu/research/tools/vocabularies/ulan/

[4] https://intavia.acdh-dev.oeaw.ac.at/

[5] https://www.wikidata.org/

[6] http://www.kulttuurisampo.fi

A main challenge in these systems is how to select and rank the interesting connecting paths. Ranking relations is discussed, e.g., in [2,13]. The methods proposed can be divided to two main categories: data-centric, where the ranking is based on properties of the graph such as frequency and specificness of a connection, and user-centric, where connections are ranked based on user given criteria.

WiSP [14] finds several paths with a relevance measure between two resources in the WikiData[7] KG, and ranks them using ranking algorithms. In [15] two algorithms and a tool RECAP are presented for explaining connections: an algorithm based on explaining individual paths between given resources in a knowledge graph, and an algorithm where additional schema information and a target predicate are used for focusing on most interesting explanations. Explanations have been studied also in the context of recommender systems [16].

Some applications, e.g., RelFinder[8] [17,18] and Explass [19], allow filtering relations between two entities with facets, the user typically has to preselect the two entities before faceted search can be used to filter the relations between the entities. In RelFinder the user selects two or more resources, and the result is a visualized graph showing how the query resources are related with each other generated by the application dynamically.

Most of the case studies in this paper apply the knowledge based method for mining relations presented in [4]. In this concept interesting relation types are defined using predefined forms and applied to an existing KG to form a new KG of interesting relations. This knowledge-based method was originally developed for the semantic portal "BiographySampo – Finnish Biographies on the Semantic Web"[9] [5,6]. The data with some modifications is now also available as part of the InTaVia KG. The idea of knowledge-based relational search turned out to be in this case feasible, and was used in developing one application perspective [4] for the in-use semantic portal BiographySampo.

For BiographySampo relations were extracted from a KG created based on KG of life events created based on biographies. Predefined SPARQL CONSTRUCT queries representing interesting connection types were used. In this way 1) non-sense connections between the query resources can be ruled out effectively by the knowledge-based rules, and 2) the explanation patterns could be used for creating natural language explanations for the connections. The question to be solved could be formulated by making selections on facets. Furthermore, the hit counts on the facets could be used to solve quantitative questions, such has "Who created most paintings depicting France". However, several challenges were encountered related to, e.g., to the explosion of the number of relations in some cases and to dealing with the direction of the semantic connections [20]. BiographySampo application limited the search to connections between persons and places, partially because it proved to be a simpler case than searching connections between persons. For example, number of interesting connections between persons can be much higher than the number of interesting connections between persons and places. Two of the three case studies presented in this paper are based on this knowledge based approach, but expand the concept to search for relations between persons.

Connection between people, and relative numbers of different types of connections, have been researched in social sciences. For example, in his famous study Mark Gra-

[7] `http://wikidata.org`

[8] `http://www.visualdataweb.org/relfinder.php`

[9] Project: `https://seco.cs.aalto.fi/projects/biografiasampo/`; portal: `https://biografiasampo.fi/`, online since 2018

novetter [21] makes a separation between “strong” and “weak” connections between persons. According to Granovetter, persons with lot of weak ties to other people tend to be the biggest innovators society. Answering these kinds of research questions requires comparing relative numbers of connections of certain type between entities.

3. Representing Relations for Faceted Search

Approach to relational search in case studies presented in this paper is founded on first extracting the interesting relations and representing the them an RDF KGs, and then querying the pre-calculated knowledge graph with SPARQL to search and analyze the relations using faceted search and various visualizations. It is possible to implement faceted search in other ways. It would be possible to, for example, dynamically search associations in a data based on initial search. However, the question of how to conceptualize the associations for search will remain the same despite the technical implementation.

Two of the case studies in this paper represent interesting relations as individual entities in a KG. To represent relations as entities the class `Relation` can be used. This is a variant of the model suggested in [4,20]. The core properties of the `Relation` class to represent directed relations from a subject resource to an object resource are:

1. type of the relation (`relationType`),
2. the subject of the relation (`relationSubject`),
3. the object of the relation (`relationObject`),
4. the explanation of the relation (`label`).

The instances of relation class have a separate property `relationType`[10] that is used to represent the nature of the relation. These relation types can include types such as “teacher of”, “collaborator of” “shared a collaborator with”. Because relations are represented as directed, separate properties for the two endpoints of the connection can be used: `relationSubject` and `relationObject`. For example, in a teacher-student relation the teacher would be represented with `relationSubject` property and the student with `relationObject` property. These would be reversed in a student-teacher relation. Note that when representing relations in this model, the relations that are not naturally directed, such as “shared teacher” connections, require two otherwise identical relation instances with separate subject and object. The `label` of the relation is a human readable explanation of the relation. In addition to these core properties the relations can have any number of other properties, such as the data source or date.

An example of a (simplified) connection instance of the class `Relation` extracted from the Getty ULAN KG is given below; it represents the patron relation between Lorenzo de Medici and Michelangelo. Here the generic relation type instantiated by a SPARQL rule is "person X was the patron of person Y".

```
[]  a                    rel:Relation ;
    rel:relationType     rel:patronOf ;            # Connection type
    rel:relationSubject  ulan:500114960 ;          # Lodenzo de Medizi
    rel:relationObject   ulan:500010654 ;          # Michelangelo
```

[10] It would be possible to treat these as sub-classes of Relation and define this as the only class of the entity, but that would be less efficient for faceted search in implementation used here.

```
    rdfs:label
        "Medici, Lorenzo de' was patron of Buonarroti, Michelangelo." .
```

The model given above represents the relations as directed relations with separate subject and object. It is also be possible to represent relations, for example, as undirected relations, or by combining them somehow as parts of larger groups, which is called "consolidated model" in this paper. Modeling solutions used here are inspired by "qualified relation" and "n-ary relation" ontology design patterns [22]. Main difference is that the nature of the relation, such as "patron-of" in the above example, is not expressed using the rdf:type property, but a separate property, in this case rel:relationType.

The example cases mainly have used a directed model because this makes detailed search possible where properties of the subject in the relation can be used in addition to the properties of the object of the relation. When the subject and object are semantically separated it is easy to search, e.g., connections between German teachers and French students. Without the semantic subject-object separation the search becomes more limited at least when implementing the search using the basic philosophy of Sampo-UI.

In the Intavia case the search is implemented only for directed relations. In the ULAN case study demonstrator was created where connections in the ULAN KG are represented in directed and undirected models and tested using a faceted search implementation. In addition to directed and undirected models, and there is also a model where relations extracted from more complex second degree connections, such as shared teacher connections, are combined under a single relation instance. The Wikidata case study on the other hand does not model relations as separate entities in KG, but ibnstead through properties.

Below is an example of a CONSTRUCT query rule used to extract `Relation` instances from the Getty ULAN KG. The query below can be used on the ULAN endpoint to get patron relations like in the example given above. The query selects an artist and a patron, and creates instances of the Relation class that have those two people as the endpoints of the directed connection: the relationSubject and the relationObject. It also creates a human readable explanation of the relation as the label of the Relation instance. The explanation is based on a simple form where names of the people in question are placed. The example given here is a minimal one.[11] The Relation instances can also include semantic information about, for example, times, and sources of the connections. For relation types that represent ternary relation, such as shared teacher relations, it might be natural to include a special property for the connecting entity.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX gvp: <http://vocab.getty.edu/ontology#>
PREFIX rel: <http://ldf.fi/schema/relations/>

CONSTRUCT {
  [] a rel:Relation ;
    rel:relationSubject ?person ;
    rel:relationObject ?person2 ;
    rdfs:label ?description ;
    rel:relationType rel:patronOf .
}
WHERE {
  # Get an artist and their patron
```

[11] You can see the queries used in ULAN case here: https://github.com/SemanticComputing/ulan-relations-conversion/tree/main/queries.

```
  ?personConcept gvp:ulan1201_patron_of ?personConcept2 .

  # Get the names
  ?personConcept a gvp:PersonConcept;
      gvp:prefLabelGVP/gvp:term ?name.
  ?personConcept2 a gvp:PersonConcept;
  gvp:prefLabelGVP/gvp:term ?name2.

  # Create a human readable explanation
  BIND (CONCAT(?name, " was patron of ", ?name2, ".") AS ?description)
}
```

Executing simple queries like the one above usually only takes a couple of seconds or less.[12] However, while queries for extracting simple first degree connections like this produce a limited number of relations, second or higher degree relations like “X and Y have a shared patron Z” can produce much larger numbers of individual relations due to combinatorial explosion. In such cases executing queries, to for example the ULAN endpoint, can take over a minute, and there can be a risk of timeout or reaching the limit of triples that the endpoint will give as result.

4. Case Studies

To demonstrate searching and analyzing relations between persons using a precalculated KG, this paper examines the following case studies: InTaVia, ULAN, and Wikipedia links. The first two cases are based on the idea of combining the knowledge-based method of extracting relation instances from an existing KG using SPARQL and using faceted search for finding and studying the instances of connections between two persons. In the Wikipedia case, links in Wikipedia articles were used and applied eith Wikidata KG. The basic idea of finding relations by applying faceted search for filtering is present in all of these cases. The created KGs are served from an Apache Jena Fuseki [13] triplestores that can be queried using SPARQL. To search, filter, and visualize the connections, web applications based on faceted search were created for each case using the Sampo-UI framework[14].

4.1. Case InTaVia: Interlinked Biographies from four European Countries

A demonstrator for relational search was created to search connections between perons in the KG created in the EU project *InTaVia: In/Tangible European Heritage*[15]. InTaVia KG combines data from four different biographical databases from Austria (APIS), Finland (BiographySampo), the Netherlands (BioraphyNet), and Slovenia (SBI), InTaVia’s mission was to transcend siloed data into a comprehensive view of European cultural heritage. The persons in the biographies are generally people considered important to the history of the respective countries. The KG has also been enriched with data from, for example, Getty ULAN and Wikidata.

[12] The above query can be be copied and executed on Getty SPARQL service `https://vocab.getty.edu/sparql`. It should take few seconds at maximum as it is a minimal example. More complex queries will be slower.

[13] `https://jena.apache.org/documentation/fuseki2/`

[14] Sampo-UI source code available on Github: `https://github.com/SemanticComputing/sampo-ui`.

[15] `https://intavia.eu/`

Relations were extracted between persons from the InTaVia data using SPARQL CONSTRUCT queries. The demonstrator where the relations can be searched is available online[16]. Currently the relations are only for Austrian, Finnish, and Slovenian persons. The demonstrator currently includes around ten thousand connections of various types including family relations and teacher-student relations. The vast majority of the connections are within a single dataset, but there are also a few hundred cases there the connections are between datasets.

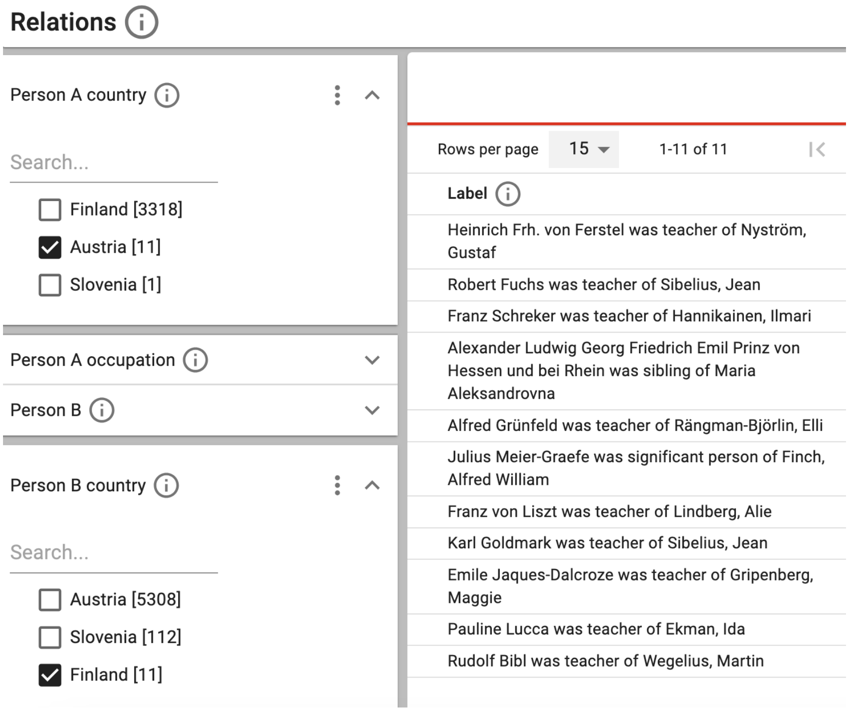
In this application the properties of the endpoints of the connection, and of the connection itself, are presented as facets. User can then make selections from the facets to narrow down the search to an interesting set of connections. Figure 1 shows an example of the user interface. The facets are located on the left side of the screen and the human readable explanations of each relation are shown on the right, as well as relevant links to the entities of the relation. The user can simply select a single entity, in this case a person, from a facet, and then look at the various relations that the selected entity has to other entities.

The user can, however, also search for relations between larger groups. For example, by making a selection from the "Occupation" facet the user is shown all relations where the person has a certain occupation. The subject and the object, called "Person A" and "Person B" in the example facets, of the relation have separate facets so that their properties can be defined separately in a search. The properties of the relation it self, mainly the type of the relation, can also have their own facets. In this case the user has searched for relations between persons in the Austrian and Finnish data sets, by making selections in the facets on the left. An example of faceted search given below in Figure 1 is a screenshot from this demonstrator. In this case the user has searched for relations where the subject in the relation is from the Austrian data and the object is from the Finnish data. Results will be then individual connections between persons in Austrian and Finnish biography collections.[17] It is easy to see that, according to the data, the Finnish composer Jean Sibelius was a student of two Austrians: Karl Goldmark and Robert Fuchs.

Such individual connections can be interesting as such, but sometimes more revealing are the hit counts on the facets on the left. In faceted search, the hit counts of facet categories tell the quantitative distributions of the results along the facet categories. The results can be ordered and visualized based on the hit count within the facet. This feature can be used for solving some quantitative research problems, in addition to finding individual relations. For example, in Figure 1 the user has searched for teacher and student relations between persons in the Finnish and Austrian data sets. It can be seen from the facets that the Austrian persons have naturally most relations to other Austrians, 5308 as shown in the lower facet. There are only 11 relations to Finland, while there is over one hundred relations to persons in the Slovenian data. Therefore there seems to be much more these kind of connections between Austria and Slovenia than between Austria and Finland. This makes sense as Slovenia is geographically and culturally much closer to Austria than Finland.

[16] `https://intaviasampo.demo.seco.cs.aalto.fi/`

[17] The actual nationality of the person is not available as a facet in the demonstrator.



Relations

Person A country

Search...

☐ Finland [3318]
☑ Austria [11]
☐ Slovenia [1]

Person A occupation

Person B

Person B country

Search...

☐ Austria [5308]
☐ Slovenia [112]
☑ Finland [11]

Rows per page 15 1-11 of 11

Label
Heinrich Frh. von Ferstel was teacher of Nyström, Gustaf
Robert Fuchs was teacher of Sibelius, Jean
Franz Schreker was teacher of Hannikainen, Ilmari
Alexander Ludwig Georg Friedrich Emil Prinz von Hessen und bei Rhein was sibling of Maria Aleksandrovna
Alfred Grünfeld was teacher of Rängman-Björlin, Elli
Julius Meier-Graefe was significant person of Finch, Alfred William
Franz von Liszt was teacher of Lindberg, Alie
Karl Goldmark was teacher of Sibelius, Jean
Emile Jaques-Dalcroze was teacher of Gripenberg, Maggie
Pauline Lucca was teacher of Ekman, Ida
Rudolf Bibl was teacher of Wegelius, Martin

Figure 1. Searching relations between persons in the Austrian and Finnish datasets of the InTaVia KG.

The relations in a result set of a search can also be visualized in different ways. For example Figure 2 shows arcs on a map[18] between birth places of women in the Austrian APIS set of biographies, and the persons they have connections. There are lot of connections inside Austria and to countries around it, whole only a few to Finland.

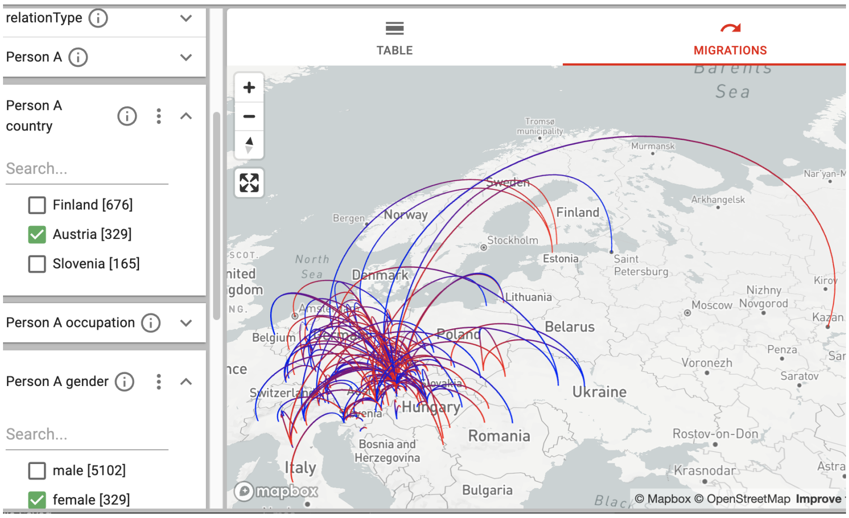
4.2. Case Relational Search in the Universal List of Artist Names (ULAN)

The Getty Universal List of Artist Names (ULAN) knowledge graph is available openly online on a SPARQL endpoint and was used as an example case[19] to test and demonstrate relational search and knowledge discovery on a well-curated dataset of artists and their mutual relations. In this case the KG includes various types of relevant connections, such as teacher-student relations or friendships between artists. Faceted search makes it possible to search not only relations between individuals, but also between larger groups, such as artists of a certain nationality or gender. In addition to the individual connections, the relative hit counts of different facet categories reveal distributions of the underlying data and can be used in explorative semantic search and browsing.

This case demonstrates also various ways to conceptualize and represent relations. In addition to the directed model presented above, this case tested an undirected model and

[18] The base map in the example is rendered by the web application using the Mapbox (`https://www.mapbox.com/about/maps/`) service which is based on the OpenStreetMap (`http://www.openstreetmap.org/copyright`.

[19] `https://www.getty.edu/research/tools/vocabularies/ulan/index.html`



relationType

Person A

Person A country

Search...

☐ Finland [676]

☑ Austria [329]

☐ Slovenia [165]

Person A occupation

Person A gender

Search...

☐ male [5102]

☑ female [329]

Figure 2. A map visualization showing arcs between birth places of women in the Austrian APIS set of biographies, and the persons they have connections.

a model where individual relations were consolidated as part of a more general relation instead of representing each one individually. For this purpose relevant data from ULAN KG was converted to a KG of relation instances using SPARQL CONSTRUCT queries[20] such as presented above, and a web demonstrator[21] was created on top of the new KG to investigate and show how faceted search works in practice using different ways of conceptualisation, and how this affects the results.

The ULAN relational search application testes and demonstrates faceted search using slightly different ways to model and conceptualize relations. Relations have been modeled as “directed”, “undirected”, and “consolidated” instances, and presented different search perspectives for each of these cases. The undirected model differs from directed in that there is only one property for both the endpoints of the connection. Therefore there is no separate subject and object in the relation. This lowers the number of required relation instances, but the numbers can still remain large. The consolidated model combines similar second degree connections under one instance. For example all the shared teacher connections stemming from a single teacher would be represented as one instance, with unlimited number of connected persons, and one connecting person (the teacher). Therefore this kind of shared teacher connections bundles all the persons that have this person as teacher under one (teacher-specific) relation. This lowers the number of required relation instances considerably, and increases performance, but the search becomes more limited, at least if the user is fundamentally interested in binary relations between two persons.

[20] The SPARQL queries used to query the ULAN KG and to create the Relation instances can be found at `https://github.com/SemanticComputing/ulan-relations-conversion`. The repository also includes code to automatically execute the queries, and a dockered Apache Fuseki server to serve the data.

[21] Demo is available at: `https://ulansampo.demo.seco.cs.aalto.fi`. The source code for the user interface is available at `https://github.com/SemanticComputing/ulan-relations-web-app`.

In practice when using faceted search through SPARQL queries the directed model has the worst performance because it has the largest number of individual relations to search, however it also was easiest to implement robust search options using SPARQL queries and the Sampo-UI framework. For example, in the demonstrator it is difficult to search relations between artists from different countries in the undirected relations perspective of the demo, while it is simple in the directed relations perspective. This is because in undirected model both entities of the relation are reached through the same property path, and therefore it is difficult to create separate facets. However, it is possible that implementing the search in different ways might make the search with other models work better.

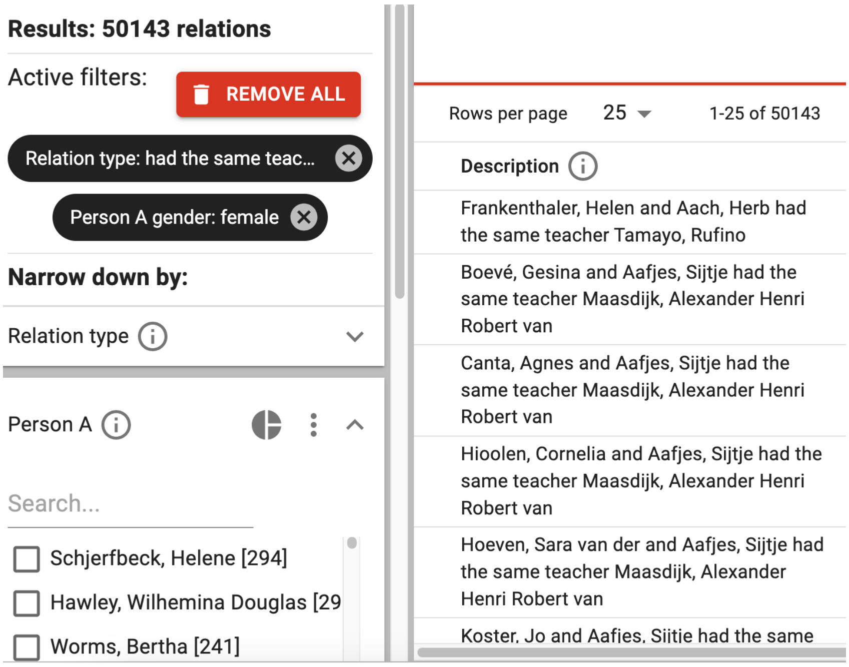
The ULAN demonstrator includes "second degree" or "ternary relations", such as shared teacher or shared patron relations. It is easy to see from the demonstrator how the the number of relations can explode for the second degree relations when using the directed relations model. While there are some tens of thousands of teacher-student relations, there are hundreds of thousands of shared teacher relations. In Figure 3 the user has selected "shared teacher" as relation type and "female" as the gender of the subject of relation. While the result set shows individual examples of connections where two artists of which at least one was female had the same teacher, the facet hit counts may be more interesting. In this case the hit count of "Person A" facet shows that the Finnish painter Helene Schjerfbeck has the largest amount of this type of relations in the data out of female painters. What this actually tells can't be answered just by looking at the demonstrator, but may well offer interesting insight to, for example, biographer of Schjerfbeck or researcher of Finnish art history.

In numerical terms the relations extracted from second degree relations such as "shared teacher" or "shared collaborator" will dwarf first degree connections such as "teacher of" or "collaborator of" relations. This quickly causes problems because the number of relations instances can grow very large. For example in the ULAN demonstrator there are more than 400 000 "shared teacher" connections, which forms a clear majority of all the connections. Such numbers start to reach limits what can efficiently searched with faceted search using only SPARQL queries. However, there can be more efficient forms of implementing faceted search that might be used on practical applications.

4.3. Case Wikipedia Links

The key idea of this case study was to harvest interesting Cultural Heritage relations between entities by using the link structure present in Wikipedia pages. Our research hypotheses was, that the textual context in which an HTML link appears can be used as an explanation for the underlying relation. In this way a new KG could be extracted to supplement relational data from the other case studies.

Register descriptions of people are often short, and an external database can provide more detailed information about their lifetime. The InTaVia KG contains also linkage to external data publications, such as the international Wikidata. Using the linkage to Wikidata allows one to also access the related Wikipedia pages written in various languages. It turned out that out of the total of 58 864 people in the InTaVia KG that have an entry in Wikidata, approximately 14 300 people have also a page in the English Wikipedia. Table 1 shows the number of Wikipedia entries of the biographical source datasets in



Results: 50143 relations

Active filters: REMOVE ALL

Relation type: had the same teac... ☒

Person A gender: female ☒

Narrow down by:

Relation type ⓘ

Person A ⓘ

Search...

□ Schjerfbeck, Helene [294]

□ Hawley, Wilhemina Douglas [29

□ Worms, Bertha [241]

Rows per page 25 ▾ 1-25 of 50143

Description ⓘ

Frankenthaler, Helen and Aach, Herb had the same teacher Tamayo, Rufino

Boevé, Gesina and Aafjes, Sijtje had the same teacher Maasdijk, Alexander Henri Robert van

Canta, Agnes and Aafjes, Sijtje had the same teacher Maasdijk, Alexander Henri Robert van

Hioolen, Cornelia and Aafjes, Sijtje had the same teacher Maasdijk, Alexander Henri Robert van

Hoeven, Sara van der and Aafjes, Sijtje had the same teacher Maasdijk, Alexander Henri Robert van

Koster, Jo and Aafjes, Sijtje had the same

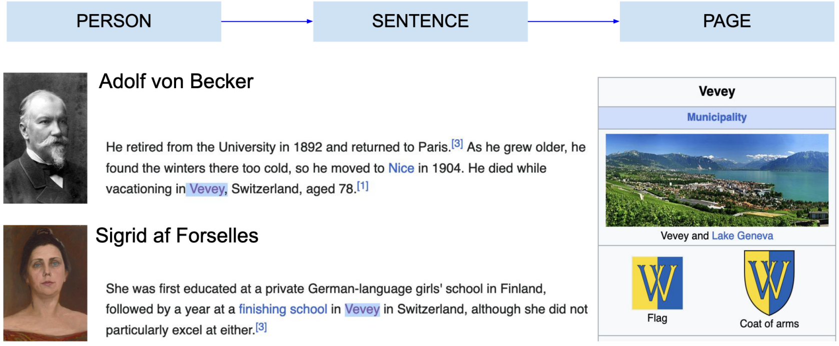
Figure 3. An example of using faceted search to explore relations in Getty ULAN KG. The facet on left shows that Finnish painter Helene Schjerfbeck has the largest number of shared teachers connections out of female artist in ULAN KG.

Table 1. Number of Wikipedia pages for the four national biographical InTaVia datasets Apis (Austria), BiographyNet (The Netherlands), BiographySampo (Finland), and SBI (Slovenia) in five different languages

Dataset	German	English	Finnish	Dutch	Slovenian
Apis	6945	3478	553	710	369
BiographyNet	5649	8022	785	14784	495
BiographySampo	1067	2180	5106	429	186
SBI	599	727	80	150	3774
Total	14260	14407	6524	16073	4824

five different languages and the four InTaVia datasets. The English pages were chosen because the number of Wikipedia matches was sufficient also for the minor languages Finnish and Slovenian.

The principle for using Wikipedia links for relational search is depicted in Figure 4. The textual description of a Person (on the left) consists of sentences (in the middle) that contain links to the pages (on the right) of new entities. In the example there are two Finnish artists, *Adolf von Becker* and *Sigrid af Forselles*, who both have a link connection to a village called *Vevey* in Switzerland. So, visiting a small village in Switzerland forms a potentially interesting relation between these two artists. After extracting all these links, they can be used as a basis for studying the network of references with explanations given by the textual contexts of the links. Furthermore, this network can be used for



He retired from the University in 1892 and returned to Paris.[3] As he grew older, he found the winters there too cold, so he moved to Nice in 1904. He died while vacationing in Vevey, Switzerland, aged 78.[1]

She was first educated at a private German-language girls' school in Finland, followed by a year at a finishing school in Vevey in Switzerland, although she did not particularly excel at either.[3]

Figure 4. Principle of modelling the Wikipedia references

prosopographical analysis to find, e.g., common features connecting two individuals, features characteristic to each source dataset, or vice versa features separating the source datasets.

In the process of data transformation from Wikipedia to a KG of explained relation instances, the description text from the Wikipedia page was first queried for each person. Thereafter, description texts were split to sentences, and the links referring to other pages in English Wikipedia were extracted. Finally, metadata of the referred entities was queried from Wikidata, most importantly the class of entity which could a person, place, work of art, genre of art or literature, etc. Based on the class of the entity, biographical or geographical details were queried and added to the data. In this process pipeline, the data was pruned by filtering out 1) links to pages that were referenced only by one single person, 2) links leading to disambiguation or multimedia pages, and 3) links leading to external web-sites.

Finally, a network was constructed based on the links in the sentences so that two people having the same link target got interconnected. The Python module WikiTextParser[22] was used for scraping the texts from the Wikipedia pages, Natural Language Toolkit (NLTK)[23] for splitting the sentences, and RDFLib[24] for producing the RDF data. As a result, 180 000 sentences referring to 37 500 Wikipedia pages were extracted with an average of 22.8 references per person.

A demonstrator portal[25] for the Wikipedia case study was also created using the Sampo-UI framework. The portal can be used to search and visualize the data with tables, charts, networks, and illustrations on maps. The portal contains three faceted application perspectives:

1. The People perspective makes it possible to find easily links to related entities on the pages of people.

[22] https://pypi.org/project/wikitextparser/

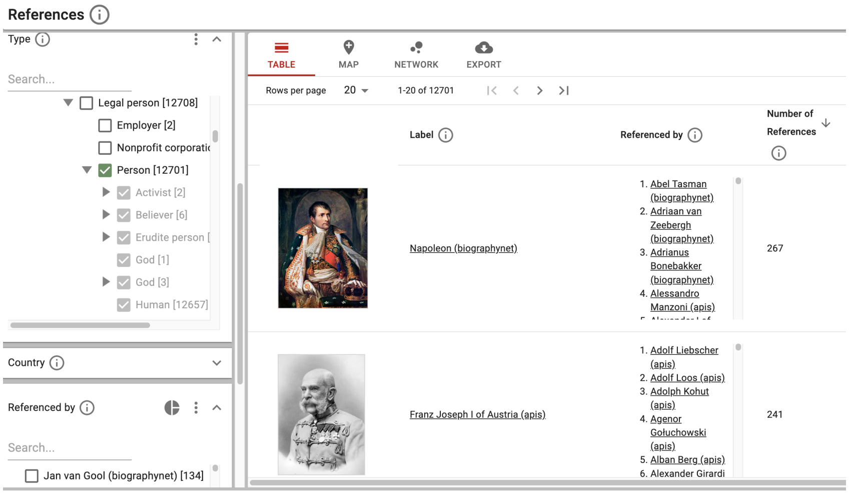
[23] https://www.nltk.org/

[24] https://pypi.org/project/rdflib/

[25] Demo is available at: https://intapedia.demo.seco.cs.aalto.fi/en/. Portal source code is available at: https://github.com/SemanticComputing/intavia-wikipedia-web-app.

2. The References perspective work the other way around: here referenced entity pages are searched for and entity pages from which links to them have been made can be found easily for each referenced entity.
3. The Sentences perspective is used for searching the sentences in which reference links occur, based on People, References, Dataset facets. The sentences explain the relation links by showing their context.

In this case study the search is directed to entities in Wikidata and Wikipedia. The search therefore behaves differently than the search perspectives in other two case studies, even though here too the focus is on finding relations. For example, looking at the References perspective it is immediately obvious, because the results are ordered based on number of references, that most referenced entities in Wikipedia biografies of persons are places. In this case Vienna and Amsterdam are the most referenced. This is not surprising because most of the biographies used are from Austria and Netherlands, although Vienna has the most references even though there are more biographies from Netherlands. Faceted search can be used to, for example, to select the "Person" as the type of entity referenced. Fig 5 shows a screenshot where user has made this selection. Result view shows that emperor Napoleon and emperor Franz Joseph I of Austria are the most referenced persons in the biographies used. Just below them are artists Rubens and Rembrandt. Finding persons that are connected to a certain person by a reference to same Wikipedia article can be done by, for example, selecting a certain person, such as Rembrant, from the "Referenced by" facet, and then looking at the other persons in the "Referenced by" column on the results.



- ☐ Legal person [12708]
  - ☐ Employer [2]
  - ☐ Nonprofit corporatio
  - ☑ Person [12701]
    - ☑ Activist [2]
    - ☑ Believer [6]
    - ☑ Erudite person [
    - ☑ God [1]
    - ☑ God [3]
    - ☑ Human [12657]

☐ Jan van Gool (biographynet) [134]

1-20 of 12701

Label	Referenced by	Number of References
Napoleon (biographynet)	1. Abel Tasman (biographynet) 2. Adriaan van Zeebergh (biographynet) 3. Adrianus Bonebakker (biographynet) 4. Alessandro Manzoni (apis)	267
Franz Joseph I of Austria (apis)	1. Adolf Liebscher (apis) 2. Adolf Loos (apis) 3. Adolph Kohut (apis) 4. Agenor Gołuchowski (apis) 5. Alban Berg (apis) 6. Alexander Girardi	241

Figure 5. An example of faceted search of Wikipedia references.

5. Discussion

Creating a precalculated KG of relations made implementing faceted search on web app easier than it would have been otherwise, and allowed the use of preexisting framework.

It would be possible to implement searching relations dynamically. However, making the transformation during a separate pre-processing phase has many benefits:

- Implementation of search on web application can be easier.
- Using a pre-compiled KG is computationally faster when querying the data.
- The pre-compiled KG can be validated and debugged more easily than corresponding dynamically run complex queries.
- Relational instances from separate, semantically incompatible KGs can be aggregated easily into the new KG.

A drawback of pre-compiling data is that the size of the transformed KG may explode in terms of the number of relations, depending on the case. Possibly some kind of hybrid solutions on pre-compiling and dynamic federated evaluation of queries could be developed. The large number number of searched instances can also make the faceted search computationally heavy and perhaps slow for the user, even though pre-compiling data does help there.

Perhaps the most important drawback of the knowledge based approach to relational search is that the interesting relation types need to be defined by humans to be reliable, at least with current technology. This can be done using SPARQL CONSTRUCT queries or some other solution, but it will in any case likely require time and technical understanding. If the exact nature of the connection is not a focus, like in the Wikipedia case presented above, the process is easier to automate.

While finding single connections and their explanations between two entities can be interesting, connections between larger categories of people can be even more interesting in revealing patterns in the data. These can be found and illustrated through faceted search and visualizations. The larger sets of connections can be seen through ontologies connected to the entities, such as occupation and place ontologies with hierarchies.

When searching connections between different types of entities like people and places, it is easy for the user to understand which properties in the faceted search are related to which entity of the connection. For example, when searching for connections between people and places, it is obvious that the occupation facet references the person in the connection. This is more complicated when both entities are of the same type, such as two people. Modeling relations as directed connections makes it possible to make clear semantic distinctions and create separate facets for both endpoints of the connection, even when they are of the same type. The user can then search for, for example, the connections between artists and writers in the KG. The drawback of this is that the connections need to be created twice so that both persons are the subject and object in one relation instance, even when the connection is fundamentally the same. This can be confusing for the user, and it creates a double the number of relation instances which slows down the faceted search.

In addition to finding and explaining interesting connections in KGs the idea of representing connection networks can be used for studying and visualizing semantic associations statistically, using timelines, on maps, and using methods of network analysis. Our examples are meant to demonstrate how relational search might be useful in CH research. It can be useful for finding individual connections and on the other hand finding larger patterns. For example, it might be interesting for an art historian to find an obscure individual connection between two artists that might then explain the professional development of one or both of the artists. On the other hand researcher might be interested in more general connections, such as the connections that Finnish female artists have to Germany in the 19th century. Such query will require filtering interesting connections from certain time period, between persons of certain gender,profession and nationality to places within certain larger place. Such a query would be difficult using traditional search methods, but KGs and ontologies can help make the query easier. A researcher might also be interested in relative numbers or connections. For example, are there more interesting connections between Finnish female artists in 19th century to Germany or to France, and how do the numbers change in time, or did Finnish artists study more often under Italian or Austrian artists? Faceted search and visualizations of relations can be useful when exploring data to answer such questions. The pre-precombiled KG of relations could also be analyzed using different methods. For example, by using graph and network analysis.

Acknowledgments Our research was supported by the EU project InTaVia. CSC – IT Center for Science, Finland, provided computational resources.

References

[1] Sheth A, Aleman-Meza B, Arpinar IB, Bertram C, Warke Y, Ramakrishnan C, et al. Semantic Association Identification and Knowledge Discovery for National Security Applications. Journal of Database Management on Database Technology. 2005;16(1):33-53.

[2] Cheng G, Shao F, Qu Y. An Empirical Evaluation of Techniques for Ranking Semantic Associations. IEEE Transactions on Knowledge and Data Engineering. 2017;29(11):1.

[3] Tunkelang D. Faceted search. Synthesis Lectures on Information Concepts, Retrieval, and Services. 2009;1(1):1-80.

[4] Hyvönen E, Rantala H. Knowledge-based Relational Search in Cultural Heritage Linked Data. Digital Scholarship in the Humanities (DSH). 2021;36:155-64.

[5] Hyvönen E, Leskinen P, Tamper M, Rantala H, Ikkala E, Tuominen J, et al. BiographySampo – Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research. In: The Semantic Web. 16th International Conference, ESWC 2019. Springer–Verlag; 2019. p. 574-89.

[6] Tamper M, Leskinen P, Hyvönen E, Valjus R, Keravuori K. Analyzing Biography Collection Historiographically as Linked Data: Case National Biography of Finland. Semantic Web – Interoperability, Usability, Applicability. 2023;14(2):385-419. Available from: `https://doi.org/10.3233/SW-222887`.

[7] Ikkala E, Hyvönen E, Rantala H, Koho M. Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. Semantic Web – Interoperability, Usability, Applicability. 2022;13(1):69-84.

[8] Rantala H, Ahola A, Ikkala E, Hyvönen E. How to create easily a data analytic semantic portal on top of a SPARQL endpoint: introducing the configurable Sampo-UI framework. In: Proceedings of 8th International Workshop on the Visualization and Interaction for Ontologies and Linked Data co-located with the 22nd International Semantic Web Conference (ISWC 2023) in Athens, Greece; 2023. .

[9] Viswanathan V, Ilango K. Ranking semantic relationships between two entities using personalization in context specification. Information Sciences. 2012;207:35-49.

[10] Hyvönen E. Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery. Semantic Web. 2020;11(1):187-93.

[11] Hyvönen E, Mäkelä E, Kauppinen T, Alm O, Kurki J, Ruotsalo T, et al. CultureSampo – Finnish Culture on the Semantic Web 2.0. Thematic Perspectives for the End-user. In: Museums and the Web 2009, Proceedings. Archives and Museum Informatics, Toronto; 2009. Available from: `https://seco.cs.aalto.fi/publications/2009/hyvonen-et-al-culsa-mw-2009.pdf`.

[12] Mäkelä E, Ruotsalo T, Hyvönen E. How to deal with massively heterogeneous cultural heritage data—lessons learned in CultureSampo. Semantic Web – Interoperability, Usability, Applicability. 2012;3(1):85-109.

[13] Bianchi F, Palmonari M, Cremaschi M, Fersini E. Actively Learning to Rank Semantic Associations for Personalized Contextual Exploration of Knowledge Graphs. In: Blomqvist E, Maynard D, Gangemi A, Hoekstra R, Hitzler P, Hartig O, editors. The Semantic Web. Cham: Springer–Verlag; 2017. p. 120-35.

[14] Tartari G, Hogan A. WiSP: Weighted Shortest Paths for RDF Graphs. In: Proceedings of VOILA 2018. CEUR Workshop Proceedings, vol. 2187; 2018. p. 37-52.

[15] Birró G. Building relatedness explanations from knowledge graphs. Semantic Web – Interoperability, Usability, Applicability. 2020;10(6):963-90.

[16] Herlocker JH, Konstan JA, Riedl J. Explaining Collaborative Filtering Recommendations. In: Computer Supported Cooperative Work. ACM; 2000. p. 241-50.

[17] Lehmann J, Schüppel J, Auer S. Discovering Unknown Connections—the DBpedia Relationship Finder. In: Proc. of the 1st Conference on Social Semantic Web (CSSW 2007). vol. 113 of LNI. GI; 2007. p. 99-110. Available from: http://subs.emis.de/LNI/Proceedings/Proceedings113/gi-proc-113-010.pdf.

[18] Lohmann S, Heim P, Stegemann T, Ziegler J. The RelFinder User Interface: Interactive Exploration of Relationships between Objects of Interest. In: Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI 2010). ACM; 2010. p. 421-2. Available from: http://doi.acm.org/10.1145/1719970.1720052.

[19] Cheng G, Zhang Y, Qu Y. Explass: exploring associations between entities via top-K ontological patterns and facets. In: International Semantic Web Conference (ISWC). Springer–Verlag; 2014. p. 422-37.

[20] Rantala H, Hyvönen E, Leskinen P. Finding and explaining relations in a biographical knowledge graph based on life events: Case BiographySampo. In: ESWC 2023 Workshops and tutorials joint proceedings. CEUR Workshop Proceedings; 2023. In press.

[21] Granovetter M. The strength of weak ties. American journal of sociology. 1973;78(6):1360-80.

[22] Dodds L, Davis I. Linked Data Patterns: A pattern catalogue for modelling, publishing, and consuming Linked Data; 2022. Available from: http://patterns.dataincubator.org/book/.