Searching interesting relations in cultural heritage knowledge graphs

Heikki Rantala[1][0000−0002−4716−6564], Petri Leskinen[1][0000−0003−2327−6942], Lilli Peura[1], and Eero Hyvönen[1,2][0000−0003−1695−5840]

[1] Semantic Computing Research Group (SeCo), Department of Computer Science, Aalto University, Finland
[2] Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland
firstname.lastname@aalto.fi

Abstract. In *relational search* interesting connections between entities in Knowledge Graphs (KG) are searched for. This paper presents an approach where potentially interesting relations are mined out in the source data and are then represented as instances in an RDF graph. These instances can then be efficiently searched using faceted search, studied, and visualized with data-analytic tools statistically, on maps, timelines, and using methods of network analysis. Four case studies are presented where this approach is applied to the BiographySampo KG, the Getty ULAN KG of artists, InTaVia KG of biographies from four European countries, and the English Wikipedia combined with Wikidata.

1 Introduction

Relational search [22], also referred as semantic association search [5], is a search paradigm where the goal is to find relations or connections between entities usually in a context of an RDF[3] knowledge graph (KG).

This paper presents a knowledge-based approach for relational search and multiple case studies of applying it to openly available RDF knowledge graphs of Cultural Heritage (CH) data. The idea of this approach is to mine potentially interesting relations in the source data and then represent them as a new RDF graph where each individual connection between two entities is represented as a separate instance. These instances of relations can be queried with SPARQL[4] and efficiently searched using, for example, faceted search [25], and studied and visualized using data-analytic methods. We present case studies where the approach is applied to the Finnish biographical BiographySampo KG, the Getty ULAN knowledge graph of artists[5], InTaVia knowledge graph of biographies from multiple European countries[6], and Wikipedia combined with Wikidata[7].

[3] https://www.w3.org/RDF/
[4] https://www.w3.org/TR/sparql11-query/
[5] https://www.getty.edu/research/tools/vocabularies/ulan/
[6] https://intavia.acdh-dev.oeaw.ac.at/
[7] https://www.wikidata.org/

2 Related Work

In relational search the *query* consists of two or more resources, and the task is to find interesting semantic associations between them. The approaches [5] 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 a few different fields. In [22] the idea of searching relations is applied for association finding in national security domain. In [27] the concept is applied to genetics and medicine research. Relational search has also been applied in the Cultural Heritage field [11,14]. CultureSampo[8] [10,20] contains an application where connections between two persons were searched using a breadth-first algorithm.

A main challenge in these systems is how to select and rank the interesting paths. Ranking relations is discussed, e.g., in [5,2]. The methods proposed include at least 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 some user given criteria.

In RelFinder[9] [17,19,8,7] the user selects two or more resources, and the result is a visualized graph showing how the query resources are related with each other. WiSP [24] finds several paths with a relevance measure between two resources in the WikiData[10] KG, using ranking algorithms. The query results are graph paths that can be ranked based on how familiar the elements related to the information are to the user [1]. Some applications, e.g., RelFinder and Explass [6], allow filtering relations between two entities with facets, but the user typically has to preselect the entities before faceted search can be used.

In [4] two algorithms and a tool RECAP are presented for explaining connections: E4D based on explaining individual paths between given resources in a knowledge graph, and E4S where additional schema information and a target predicate are used for focusing on more interesting explanations. In contrast to these, our method is not based on the schema but on additional domain knowledge patterns of interestingness, that are used both for finding the connecting paths in the first place, and for explaining them. Explanations have been studied also in the context of recommender systems [9].

3 Representing Relations in a Knowledge Graph

Our approach to relational search is founded on first representing the interesting relations as individual entities in an RDF KGs and then querying the knowledge graph to search and analyze the relations using faceted search and various visualizations. In most of our examples we have used a knowledge based method [14,21] where we have mined existing KGs using predefined SPARQL CONSTRUCT forms that represent potentially interesting relations to create the

[8] `http://www.kulttuurisampo.fi`
[9] `http://www.visualdataweb.org/relfinder.php`
[10] `http://wikidata.org`

new KG with relation instances. However the approach is not limited to only to existing KGs. In our Wikepedia/Wikidata example we have relied largely to text in wikipedia articles to mine relations for a KG.

3.1 Transforming the Knowledge Graph

Firstly, a data model for the transformed KG is needed. In the most simple case we use class `Relation` as the class for the relations. We use a separate property `relationType` to represent the nature of the relation. These relation types can include for example "teacher of" or "collaborator of" types of relations. It is possible to represent relations using undirected or directed models, but we have used a directed model because that makes more detailed search possible. For example, it is then possible to search teachers and students separately from teacher-student relations. Therefore there are separate properties for the two endpoints of the connection: `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. The `label` of the relation is a human readable explanation of the relation. In addition to these core properties the relations can have other properties such as the data source or date. The core properties of the `Relation` class to represent directed relations are then:

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`).

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

Below is an example of a CONSTRUCT QUERY rule used to create `Relation` instances. In this example relations are mined 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 slightly simplified and minimal one. The Relation instances can also include semantic information about, for example, times, and sources of the connections. The queries like this are not too computationally demanding, and executing queries like the one below usually only takes a couple of seconds. However, while queries mining for simple first degree connections like this produce a limited number of relations, second or higher degree relations like shared teacher or shared patron can produce much larger numbers of individual relations due to combinational explosion. In such cases there can be a risk of timeout or reaching the limit of triples that the endpoint will give as result.

```
PREFIX re: <http://www.w3.org/2000/10/swap/reason#>
PREFIX skosxl: <http://www.w3.org/2008/05/skos-xl#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
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
  ?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)
}
```

The model given above represents the relations as directed relations with separate subject and object. This is not the only possibility. It would also be possible to represent relations also, for example, as undirected relations, or by combining them somehow as parts of larger groups. The obvious benefit of undirected model would be that, for example, such “naturally undirected” relations as shared-teacher connection would need only one relation instance, while the directed model requires two. On the other hand for “naturally directed relations”, such as teacher-student relation, the undirected model would lose the semantic information of who is the teacher and who is the student in the relation. Generally, at least using our tools, it is easier to implement a more full faceted search for directed relations although the performance can suffer because of larger number of entities.

4 Case Studies

To test and demonstrate the knowledge-based method above, four case studies have been conducted and are introduced below. BiographySampo, InTaVia, and ULAN cases are all based on the idea of combining knowledge based method of extracting relation entities presented above with faceted search. The Wikipedia case differs in that Wikipedia text is used to find connections in addition to the Wikidata KG. Because the Wikipedia case is somewhat different it is discussed more broadly in a later section. Basic ideas of representing the relations as instances in a KG and applying faceted search for filtering are present in all of these cases. In all cases the KG is served from an Apache Jena Fuseki [11] triplestore that can be queried using SPARQL.

4.1 Case BiographySampo: Biographies on the Semantic Web

The knowled based method outlined above was originally developed for the in-use semantic portal "BiographySampo – Finnish Biographies on the Semantic Web"[12] [13,23]. BiographySampo publishes biographical data included in the National Biography of the Finnish Literature Society[13] about historical Finnish persons, and is part of the Sampo series of LOD services[14] and portals [12]. The contents are expressed as Linked Open Data (LOD) using the event-based Bio CRM [26] model designed for biographical data, an extension of CIDOC CRM[15]. The data is now also available as part of the InTaVia KG.

The biographical data of BiographySampo is based on 13 144 biographies and includes 266 340 life events, including births, deaths, career events, received accolades, and even historical events where the persons have participated in. The life of each biographee is described semantically in terms of spatio-temporal events which they participated in. The event data was extracted from the semi-structured summaries included in the biographies using regular expressions. [23] BiographySampo has also been enriched from other data sources, such as the HISTO[16] ontology of Finnish historical events, open data of the Finnish National Gallery, other Sampo systems, and Wikidata. These events can create various types of connections between persons and places. For example two people might have been born in the same place or participated in the same historical event.

The idea of knowledge-based relational search turned out to be in this case feasible and was used in developing one application perspective [14]for the in-use semantic portal BiographySampo. However, also several challenges were encountered related to, e.g., to the explosion of the number of relations in some cases

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

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

[13] National Biography of Finland: `https://kansallisbiografia.fi/`

[14] Information about the Sampo systems can be found here: `https://seco.cs.aalto.fi/applications/sampo/`

[15] `http://cidoc-crm.org`

[16] `https://seco.cs.aalto.fi/ontologies/histo/`

and to dealing with the direction of the semantic connections [21]. To investigate these issues of the method further, the case studies below were conducted as part of the InTaVia project using different kind of KGs.

4.2 Case InTaVia: Biographies from Europe

We have created another demonstrator as part of the EU project InTaVia: In-/Tangible European Heritage[17]. InTaVia KG comnbines 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. As part of the InTaVia project we mined relations between persons from the InTaVia data. The demonstrator where the relations can be searched is available online[18] and the source code can be examined at Github[19]. The example of faceted search given below in Fig. 1, is a screenshot from this web application. 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. Vast majority of the connections are within a single dataset. There are a few hundred cases there the connections are between datasets.

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

The Getty Universal List of Artist Names (ULAN) knowledge graph available openly online on a SPARQL endpoint was used as an example case[20] to test and demonstrate relational search and knowledge discovery on a well-curated dataset of artists and their 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 selections reveal statistical distributions of the underlying data and can be used in focusing explorative semantic search and browsing.

In this case we have tested various ways to conceptualize and represent relations. In addition to the directed model presented above, we tested an undirected model an a model where individual relations were consolidated as part of a more

[17] `https://intavia.eu/`
[18] `https://intaviasampo.demo.seco.cs.aalto.fi/`
[19] `https://github.com/SemanticComputing/intaviasampo-web-app`
[20] `https://www.getty.edu/research/tools/vocabularies/ulan/index.html`

general relation instead of representing each one individually. We have transformed[21] relevant data from ULAN KG to a KG of relation instances, and created a demonstrator *ULANSampo*[22] on top of the new KG to show how faceted search works in practice using different ways of conceptualisation, and how this affects the results.

5 Finding and Explaining Semantic Relations by Faceted Search

To search, filter, and visualize the connections, we use a web application based on faceted search. The `Relation` instances and the ontologies relating to the people and places are served from an RDF triple store, and queried by the application using SPARQL. We have published a demo for searching connections between people and places as part of the BiographySampo portal[23]. This application was implemented using the SPARQL Faceter tool [16], and is partially documented in [14]. For the more recent example cases discussed here we have created new web applications with expanded functions that are based on the new, more versatile Sampo-UI framework [15,21]. The example below is based on the demo of the InTaVia KG-based system that is available online[24]. 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. Based on the results right we can see, for example, that Austrian composer Karl Goldmark was a teacher of the Finnish composer Jean Sibelius.

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

[21] The code and the queries used can be accessed at `https://github.com/SemanticComputing/ulan-relations-conversion`.

[22] Demo available at: `https://ulansampo.demo.seco.cs.aalto.fi`

[23] `http://biografiasampo.fi/yhteyshaku/`

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

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. The relations in a result set of a search can also be visualized in different ways, such as with timelines or on maps as shown later when discussing the Wikipedia case study.

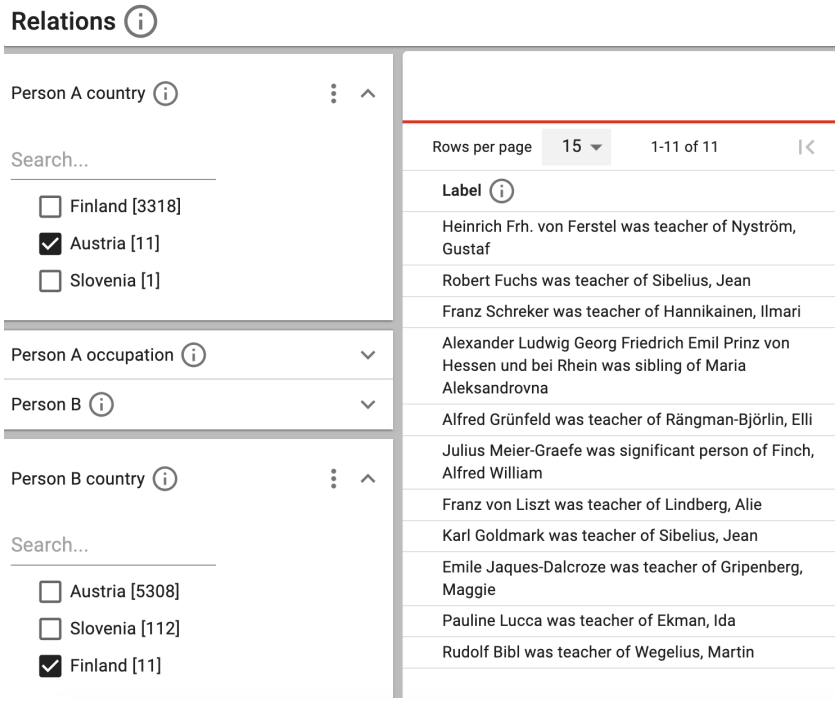

Fig. 1: Searching relations between persons in the Austrian and Finnish datasets of the InTaVia KG.

In the ULAN relational search application[25] we have tested using slightly different ways to model relations. We have modeled relations as "directed", "undirected", and "consolidated" instances, and presented different search per-

[25] `https://ulansampo.demo.seco.cs.aalto.fi`

spectives for each of these cases. 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 offers the most robust search options. For example, 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.

The ULAN demonstrator includes also second degree relations, such as shared teacher or shared patron relations. It is easy to see from the demonstrator how the the number of of relations can explode for the second degree relations. While there are some tens of thousands of teacher-student relations, there are hundreds of thousands of shared teacher relations.

6 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 set of relation instances with explanations could be created and added into the KG to supplement relational data from the other case studies.

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

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 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.

Fig. 2: Principle of modelling the Wikipedia references

The principle for using Wikipedia links for relational search is depicted in Figure 2. 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 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.

KG transformation 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[26] was used for scraping the texts from the Wikipedia pages, Natural Language Toolkit (NLTK)[27] for splitting the sentences, and RDFLib[28] 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 two-dimensional embedding of the data is depicted in Figure 3 where the datapoints representing people are colorcoded by the corresponding InTaVia dataset. The Dutch BiographyNet and the Austrian Apis are the most dominant datasets. On the other hand, a large portion of actors in the Finnish BiographySampo remain in a few separate clusters. Likewise, the largest clusters of the Slovenian SBI are located close to datapoints belonging to Apis.

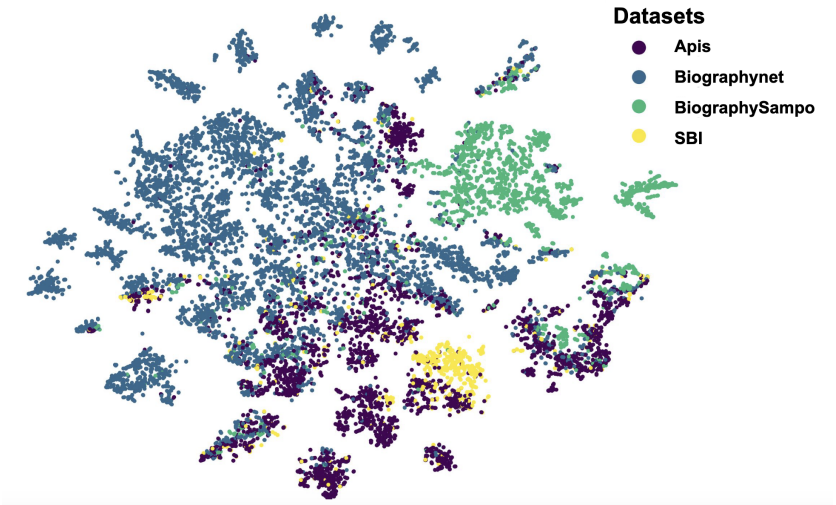

Fig. 3: Clusters of National Datasets in the InTaVia KG based on linkage in the English Wikipedia

Analysis By the nature of creating and extracting the data, the most renowned actors tend to get the largest number of references. Naturally, a famous person has a longer description text containing also a larger amount of links. Table 2 lists actors with the ten highest network centrality values. The first column *indegree* stands for the number of links pointing to that particular actor, while *outdegree* is the number of links from her or his page in Wikipedia. *Pagerank* in the third column is one of the most used centrality measures developed originally by Google [3].

[26] `https://pypi.org/project/wikitextparser/`
[27] `https://www.nltk.org/`
[28] `https://pypi.org/project/rdflib/`

The fourfold table in Figure 4 shows a mapping of terms related to visual arts. Here we have selected only such terms that are referenced in all four datasets. The terms are positioned by their proportional frequencies in the InTaVia datasets so that the terms that are equally referenced in all datasets appear in the middle of the figure like *Realism*, *Impressionism*, or *Naturalism*. On the other hand, terms that are more specific to one of the datasets appear at the corners of the grid. For example, *Surrealism* and *Art of sculpture* at the left upper corner are most commonly referenced in BiographyNet entries, *Functionalism* at right upper corner in BiographySampo, and *Symbolism* at right lower corner in Apis.

Table 2: Most central actors in the Wikipedia link network ordered by their three centrality measures

	Indegree	Outdegree	Pagerank
1	Napoleon (bnet)	Jan van Gool (bnet)	Napoleon (bnet)
2	Franz Joseph I of Austria (apis)	David Teniers the Younger (bnet)	Franz Joseph I of Austria (apis)
3	Peter Paul Rubens (bnet)	William the Silent (bnet)	Adolf Hitler (apis)
4	Rembrandt (bnet)	Christina of Sweden (bs)	Charles V (bnet)
5	Charles V (bnet)	Joost van den Vondel (bnet)	Ludwig van Beethoven (apis)
6	Ludwig van Beethoven (apis)	Peter Paul Rubens (bnet)	Louis XIV of France (bnet)
7	William the Silent (bnet)	Rembrandt (bnet)	Philip II of Spain (bnet)
8	William III of England (bnet)	Cornelis Hofstede de Groot (bnet)	Peter Paul Rubens (bnet)
9	Louis XIV of France (bnet)	Andries de Graeff (bnet)	William the Silent (bnet)
10	Arnold Houbraken (bnet)	Jan van de Cappelle (bnet)	Rembrandt (bnet)

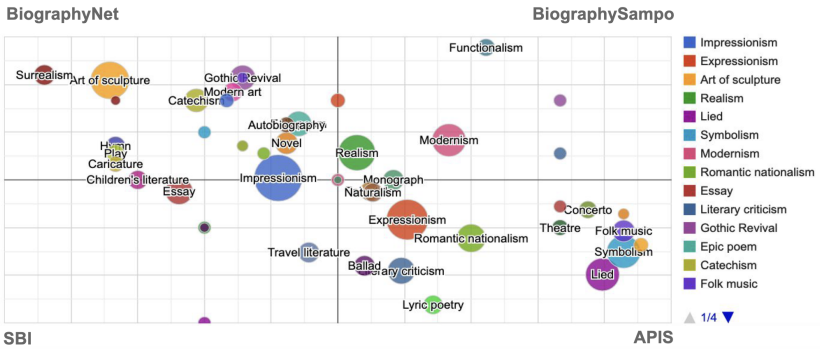

Fig. 4: Fourfold Table of Terms related to Art in InTaVia Datasets

Portal demo A portal demostrator[29] for the Wikipedia case study was created based on the Sampo–UI framework containing faceted search perspectives for the actors, the sentences in the biographical descriptions, and referenced entities. In the portal, the data can be visualized using data tables, charts, networks, and illustrations on maps. Figure 5 depicts a time series of yearly births and deaths. There are three distinguishable focal periods in the figure. Firstly, the Dutch Golden Age 1575–1675 is emphasized due to the large number of people provided by BiographyNet. Secondly, the people of late 19th century who are well represented in all four datasets. Finally, the third peak takes place at the last years of World War II due to the large number of casualties in BiographyNet, Apis, and SBI.

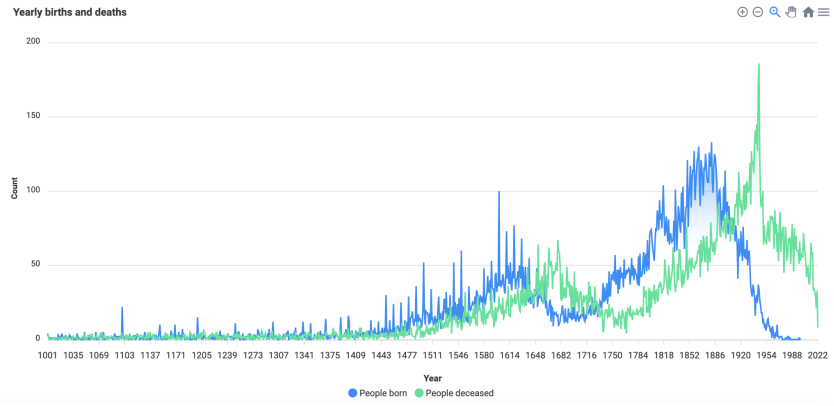

Fig. 5: Time series depicting the yearly births and deaths

Figure 6 shows two views of the map application perspective of the portal demonstrator. The one on the left is a heatmap visualization with the hotspots at Amsterdam and at Vienna due to the national biographies of the Netherlands and Austria. The two other national major cities Helsinki and Ljubljana do not pop out in the visualization as well although they have a large amount of events in the corresponding national datasets. The figure on the right depicts the high density of the markers located at the Center of Amsterdam. Altogether there are approx. 3000 events that took place at this area.

Future work Analysis using the same concept of enriching the data by Wikipedia descriptions has earlier been studied with the project AcademySampo [18]. Here the English Wikipedia pages were used which could be extended to utilize also pages in other languages as well. The links in this preliminary study were the ones pointing from an actor's page to a references entity.

[29] `https://intapedia.demo.seco.cs.aalto.fi/en/`

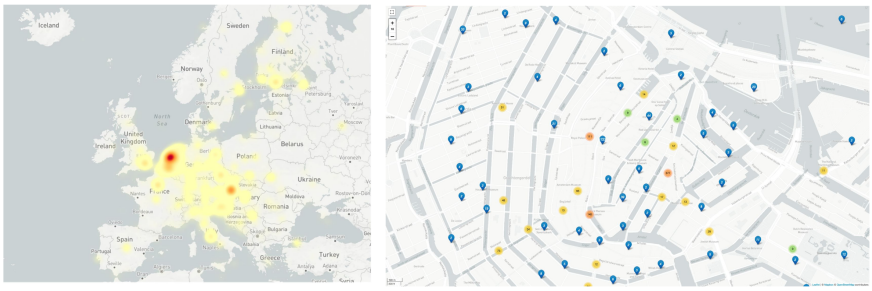

Fig. 6: Left: heatmap depicting the number of activities. Right: Map markers for activities at the Center of Amsterdam

However, the Python module using the Wikipedia API also supports querying all the links pointing to a particular page.

7 Discussion

The queries characterize different general patterns of knowledge that are likely to be interesting, and extract corresponding relational instances from the KG into a new KG. It would be possible to try to find the relations dynamically using queries when they come in. However, making the transformation during a separate pre-processing phase has many benefits:

- Using a pre-compiled KG is computationally much 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, too. We have wanted to apply faceted search for finding interesting relations. Faceted search can be resource intensive and there pre-combiled relations are useful. The data model we use is also optimized for faceted search.

While finding single connections and their explanations between two entities are interesting, also connections between larger categories of people and geographical areas are of interest and 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. We used only connections between people and places in our

earlier demonstrator [14], because this is a simple case to start with. Working on connections between persons offers new challenges. The challenges addressed in this paper are twofold: 1) How to define relations so that their number stays manageable for faceted search, which can be resource consuming if the number of searched entities is large. 2) How to implement the faceted search user interface.

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. The reason we use directed connections is that it makes it possible to 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. relational search 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.

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

References

1. Al-Tawil, M., Dimitrova, V., Thakker, D.: Using knowledge anchors to facilitate user exploration of data graphs. Semantic Web **11**(2), 205–234 (2020). https://-

doi.org/10.3233/SW-190347
2. 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. (eds.) The Semantic Web. pp. 120–135. Springer–Verlag, Cham (2017). https://-doi.org/10.1007/978-3-319-58068-5_8
3. Bianchini, M., Gori, M., Scarselli, F.: Inside PageRank. ACM Transactions on Internet Technology **5**(1), 92–128 (2 2005). https://doi.org/10.1145/1052934.1052938
4. Birró, G.: Building relatedness explanations from knowledge graphs. Semantic Web – Interoperability, Usability, Applicability **10**(6), 963–990 (2020)
5. Cheng, G., Shao, F., Qu, Y.: An empirical evaluation of techniques for ranking semantic associations. IEEE Transactions on Knowledge and Data Engineering **29**(11), 1 (2017)
6. Cheng, G., Zhang, Y., Qu, Y.: Explass: exploring associations between entities via top-k ontological patterns and facets. In: International Semantic Web Conference (ISWC). pp. 422–437. Springer–Verlag (2014)
7. Heim, P., Hellmann, S., Lehmann, J., Lohmann, S., Stegemann, T.: Relfinder: Revealing relationships in rdf knowledge bases. In: Proceedings of the 4th International Conference on Semantic and Digital Media Technologies (SAMT 2009). pp. 182–187. Springer–Verlag (2009), http://dx.doi.org/10.1007/978-3-642-10543-2_21
8. Heim, P., Lohmann, S., Stegemann, T.: Interactive relationship discovery via the semantic web. In: Proceedings of the 7th Extended Semantic Web Conference (ESWC 2010). vol. 6088, pp. 303–317. Springer–Verlag, Berlin/Heidelberg (2010), http://dx.doi.org/10.1007/978-3-642-13486-9_21
9. Herlocker, J.H., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Computer Supported Cooperative Work. pp. 241–250. ACM (2000)
10. Hyvönen, E., Mäkelä, E., Kauppinen, T., Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Takala, J., Puputti, K., Kuittinen, H., Viljanen, K., Tuominen, J., Palonen, T., Frosterus, M., Sinkkilä, R., Paakkarinen, P., Laitio, J., Nyberg, K.: 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), https://seco.cs.aalto.fi/publications/2009/hyvonen-et-al-culsa-mw-2009.pdf
11. Hyvönen, E.: Using the semantic web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. Semantic Web **11**(1), 187–193 (2020). https://doi.org/10.3233/SW-190386
12. Hyvönen, E.: Digital humanities on the Semantic Web: Sampo model and portal series. Semantic Web – Interoperability, Usability, Applicability **14**(4), 729–744 (2022). https://doi.org/10.3233/SW-223034
13. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., Keravuori, K.: BiographySampo - publishing and enriching biographies on the semantic web for digital humanities research. In: Proceedings of the 16th Extendwed Semantic Web Conference (ESWC 2019). Springer–Verlag (2019)
14. Hyvönen, E., Rantala, H.: Knowledge-based relational search in cultural heritage linked data. Digital Scholarship in the Humanities (DSH) **36**, 155–164 (2021). https://doi.org/https://doi.org/10.1093/llc/fqab042
15. 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 **13**(1), 69–84 (2022)

16. Koho, M., Heino, E., Hyvönen, E.: 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. pp. 53–63. CEUR Workshop Proceedings (2016), http://ceur-ws.org/Vol-2187/paper5.pdf
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). LNI, vol. 113, pp. 99–110. GI (2007), http://subs.emis.de/LNI/Proceedings/Proceedings113/gi-proc-113-010.pdf
18. Leskinen, P., Hyvönen, E.: Biographical and prosopographical analyses of finnish academic people 1640–1899 based on linked open data. In: Biographical Data in a Digital World 2022 (BD 2022), Tokyo. Proceedings, accepted (August 2023), forth-coming
19. 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). pp. 421–422. ACM (2010), http://doi.acm.org/10.1145/1719970.1720052
20. Mäkelä, E., Ruotsalo, T., Hyvönen, E.: How to deal with massively heterogeneous cultural heritage data—lessons learned in CultureSampo. Semantic Web – Interoperability, Usability, Applicability **3**(1), 85–109 (2012)
21. 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
22. Sheth, A., Aleman-Meza, B., Arpinar, I.B., Bertram, C., Warke, Y., Ramakrishnan, C., Halaschek, C., Anyanwu, K., Avant, D., Arpinar, F.S., Kochut, K.: Semantic association identification and knowledge discovery for national security applications. Journal of Database Management on Database Technology **16**(1), 33–53 (2005)
23. 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 **14**(2), 385–419 (2023), https://doi.org/10.3233/SW-222887
24. Tartari, G., Hogan, A.: WiSP: Weighted shortest paths for RDF graphs. In: Proceedings of VOILA 2018. pp. 37–52. CEUR Workshop Proceedings, vol. 2187 (2018)
25. Tunkelang, D.: Faceted search. Synthesis Lectures on Information Concepts, Retrieval, and Services **1**(1), 1–80 (2009)
26. Tuominen, J., Hyvönen, E., Leskinen, P.: Bio CRM: A data model for representing biographical data for prosopographical research. In: BD2017 Biographical Data in a Digital World 2017, Proceedings. pp. 59–66. CEUR WS Proceedings (2018), http://ceur-ws.org/Vol-2119/paper10.pdf
27. Viswanathan, V., Ilango, K.: Ranking semantic relationships between two entities using personalization in context specification. Information Sciences **207**, 35–49 (2012)