Who is Related to What and How? Using Biographical Knowledge Graphs for Explainable Relational Search in BiographySampo

Heikki Rantala[1[0000−0002−4716−6564]*] and Eero Hyvönen[1,2[0000−0003−1695−5840]]

[1] Semantic Computing Research Group (SeCo)
Aalto University and University of Helsinki, Finland
https://seco.cs.aalto.fi, firstname.lastname@aalto.fi
[2] Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland

Abstract. This paper presents a knowledge-based approach and an in-use application for finding and explaining "interesting", or even serendipitous, semantic relations between resources in a knowledge graph. The idea is to characterize the notion of interesting connection in terms of generic ontological explanation patterns that are applied to an underlying linked data repository to instantiate connections. In this way, 1) semantically uninteresting connections can be ruled out effectively, and 2) natural language explanations about the connections can be created. The idea has been implemented and tested based on a knowledge graph of biographical data extracted from the short biographies of 13 100 prominent historical persons in Finland, enriched by data linking to collection databases of museums, libraries, and archives. The demonstrator is in use as part of the BiographySampo portal of interlinked biographies that has had some 126 000 users. BiographySampo is based on an online Linked Open Data service, including a SPARQL endpoint.

Keywords: linked data, knowledge discovery

1 Knowledge Discovery as Relational Search

Research Problems This paper addresses the following problem of knowledge discovery [25] in Cultural Heritage (CH) [13] knowledge graphs (KG) [8]: *How are two concepts related to each other?* Semantic connections in a KG can be found between individual entities (e.g., how is Vincent van Gogh related to the village of Auvers-sur-Oise or to Paul Gaguin?) but also between more general concepts (e.g., how are Dutch impressionists related to France?). Such semantic connections can be based on various criteria for the underlying connecting paths. The problem of finding semantic connections has been called in semantic web research as association finding [28] or as relational search [24, 11, 10, 31].

We address the following key challenges involved in solving relational search problems from an end user perspective:

1. *How to disambiguate "interesting" [29] or even "serendipitous"*[3] *[20] semantic connections from non-interesting ones.* Concepts in a KG are related to each other in many ways, but only few of them are of interest to the user. For example, that van Gogh and Gauguin are instances of the class `owl:Class` is not interesting. Serendipitous knowledge discovery has been coined as one of the grand promises of the Semantic Web in Digital Humanities [15].
2. *How to explain a semantic connection to the end user?* Finding out an interesting connection is not enough (cf. the examples above) if the system cannot explain to the end user why the connection could be interesting. This problem is addressed, eg., in the field of explainable AI [7, 22].
   In our approach we precalculate connections between two entities, in our example people and places, based on predefined forms that represent connection types that are deemed interesting using SPARQL CONSTRUCT queries. These predefined connections, and their explanations can then be explored using faceted search, based on hierarchical ontologies that represent the properties of the entities. This allows for finding serendipitous connections between single entities through an exploratory process, but also importantly finding connections between larger groups of entities.
3. *How to formulate the query and query results when searching for connections.*

Given the richness of possible semantic connections, solving relational search problems can be seen as an instance of computational creativity [4], an example of the subtype "exploratory creativity", where creativity refers to search within a predefined search space under given constraints for the solutions.

Related Works In relational search the *query* consists of two or more resources, and the task is to find semantic relations between them. The approaches [5] differ in terms of the query formulation, underlying KG, methods for finding connections, and representation of the results. In [28] the idea of searching relations is applied for association finding in national security domain. CultureSampo[4] [14, 26] contains an application where connections between two persons were searched using a breadth-first algorithm, and the result was a list of chains of arcs (such as *student-of*, *patron-of*, etc.), connecting the persons. In RelFinder[5] [23, 24, 11, 10] the user selects two or more resources, and the result is a minimal visualized graph showing how the query resources are related with each other. In WiSP [31], several paths with a relevance measure between two resources in the WikiData KG[6] can be found, based on different weighed shortest path 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. A main challenge in these systems is how to select and

[3] Serendipity means 'happy accident' or 'pleasant surprise', even 'fortunate mistake'. According to the Merriam-Webster dictionaryserendipity is "the faculty or phenomenon of finding valuable or agreeable things not sought for".

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

[5] http://www.visualdataweb.org/relfinder.php

[6] http://wikidata.org

rank the interesting paths. This problem can be approached by focusing only on "simple paths" that do not repeat nodes, on only restricted node and arc types in the graph (e.g., social connections between persons), and by assuming that shorter, possibly weighted paths are more interesting than longer ones. For weighting paths, measures such as page rank of nodes and commonness of arcs, can be used. Ranking relations is discussed, e.g., in [5, 2].

In [3] 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 patters 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 [12].

Paper Outline This paper presents and applies a knowledge-based approach to the research problems above and, in particular, presents the in-use application Faceted Relator where the method has been tested and evaluated as part of the larger application "BiographySampo – Finnish Biographies on the Semantic Web"[7] [16, 30] that has had 126 000 users on the Web. This paper extends and complements our earlier papers [17, 18] by presenting a more technical and end user centric account of the application as well as first evaluation results.

In the following, our novel knowledge-based approach to relational search and its implementation are first presented (Section 2). After this, using the application is explained and evaluation results are presented (Section 3). In conclusion, lessons learned are discussed and further research suggested.

2 Finding Semantic Relations

Knowledge-based Method The graph-based methods above make use of generic graph traversal algorithms that are application domain agnostic. In contrast, this paper suggests a *knowledge-based* approach where the problem of relational search is reduced into a search problem on explained connections in a simpler search space that is transformed from the original KG using knowledge-based SPARQL CONSTRUCT query rules. The re-formulated search problem is then solved effectively as a faceted search problem [18] re-using a ready-to-use tool [21] for the purpose. 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 can be used for creating natural language explanations for the connections. The price to be paid is the need for crafting the transformation rules and their explanation patterns manually, based on application domain knowledge, as customary in knowledge-based system.

In the following, the original datasets used in the transformation are first presented. After this the data model of the final application is explained, and

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

how the data transformation into it was done. Finally the LOD service underlying the in-use systems is discussed.

Datasets The knowledge graph underlying our system system was created using the following interlinked datasets of BiographySampo:

1. The biographical data of BiographySampo based on 13 144 Finnish biographies in Bio CRM form [32] including, e.g., 51 937 family relations, 4953 places, 3101 occupational titles, and 2938 organizations.
2. HISTO ontology[8] of Finnish history including more than thousand historical events with related people, places, and times.
3. The Fennica National Bibliography[9], a LOD database of Finnish publications since 1488.
4. BookSampo[10] linked data covering virtually all Finnish fiction literature, maintained by the Finnish Public Libraries.
5. The Finnish National Gallery collections dataset[11] described using Dublin Core, JSON, and XML formats that was transformed into RDF.
6. The collected works of the J. V. Snellman[12], the national philosopher of Finland, with, e.g., 1500 letters.

Datamodel The key class in the new search space is Relation with the core properties listed in Table 1. The key elements there are a set of properties that explicate the resources that are connected and a literal natural language expression that explains the connection in a human readable form. In the application, we decided to search for relations between people and places and therefore used the person and place ontologies of BiographySampo as the basis of facet ontologies. The occupation ontology and place hierarchy of Biographysampo were used to allow faceted search based on properties of the entities. In addition, a new facet ontology of relation types was created.

Table 1. Metadata schema for semantic connections (class Connection) with explanations (prefLabel)

Element URL	C	Range	Meaning of the value
skos:prefLabel	1	xsd:string	Human readable explanation
:relationType	1	:Relation	Type of the relation
:personSubject	1	:Person	Subject of the relation (person)
:placeObject	1	:Place	Object of the relation (place)
:date	0..1	xsd:date	Time associated with the relation
:source	1..n	URI	Resource that is the source of the relation, such as an Event
:sourceName	1	xsd:string	Human readable description of the source for the relation

[8] https://seco.cs.aalto.fi/ontologies/histo/
[9] https://www.kansalliskirjasto.fi/en/services/conversion-and-transmission-services-of-metadata/open-data
[10] https://www.ldf.fi/dataset/kirjasampo/index.html
[11] https://www.kansallisgalleria.fi/en/avoin-data/
[12] http://snellman.kootutteokset.fi/

For example, the following illustrative example of a tertiary relation $< X, Y, Z >$ in RDF Turtle notation[13] connects the person `Leonardo_da_Vince` to the place `Vince` and the time `1452` based on the explanation "Person X was born in place Y in Z". The instance `:c123` is an individual of the connection type (class) `:BirthConnection` for birth connections that defines the generic explanation pattern for its instances using the property `rdfs:label`:

```
:c123 a :BirthConnection;
  :explanation "Leonardo da Vinci was born in Vince in 1452";
  :place :vince;
  :time 1452;
  :person :Leonardo_da_Vince .

:BirthConnection rdfs:label "Person X was born in place Y in time Z" .
```

Data Transformation The graph transformations into the data model above were performed using SPARQL[14] CONSTRUCT queries. The queries transformed (part of) the BiographySampo KG into a new KG of connection instances.

The focus in our demonstrator is on finding relations describing connections between people and places in Finnish cultural history. The relation instances listed in Table 2 were created using the SPARQL CONSTRUCT queries whose application to the data generated connection instances with related natural language explanations. For example, the following query can be used to create connections between people and their death places and times:

```
# Namespace definitions

BASE <http://ldf.fi/relse/> # Namespace for connection instances

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX skosxl: <http://www.w3.org/2008/05/skos-xl#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX crm: <http://www.cidoc-crm.org/cidoc-crm/>
PREFIX gvp: <http://vocab.getty.edu/ontology#>
PREFIX schema: <http://schema.org/>
PREFIX rel: <http://ldf.fi/relse/>
PREFIX nbf: <http://ldf.fi/nbf/>

# Template for constructing connection instances
CONSTRUCT {
    ?uri a rel:Relation ;
        rel:relationType rel:deathPlace ;
        rel:personSubject ?person ;
        rel:placeObject ?place ;
        rel:date ?deathtime ;
        rel:source ?death ;
        rel:sourceName "Tapahtuma Semanttisessa kansallisbiografiassa" ;
        skos:prefLabel ?description .
}

# Matching the variables for constructing the connections above
WHERE {
# Person
    ?death crm:P100_was_death_of/^foaf:focus ?person .
    ?person skosxl:prefLabel/schema:familyName ?familyName .
```

[13] https://www.w3.org/TR/turtle/

[14] https://www.w3.org/TR/sparql11-overview/

```
    ?person skosxl:prefLabel/schema:givenName ?givenName .
# Place
    ?death nbf:place ?place .
    ?place skos:prefLabel ?placeName .
   FILTER(lang(?placeName) = 'fi') .
# Time
    ?death nbf:time/gvp:estStart ?deathtime .
    BIND (year(xsd:date(?deathtime)) as ?year)
# URI
    BIND(uri(encode_for_uri(concat(str(?person), str(?place),
        "death_place", str(?death)))) as ?uri) .
# Natural language explanation
    BIND(concat(str(?givenName), " ", str(?familyName), " on kuollut paikassa ",
        str(?placeName), " vuonna ", str(?year), ".") as ?description) .
}
```

The query consists of the following parts marked by comment lines beginning with '#': First, the prefixes for namespaces are introduced: `xsd`, `skos`, `skosxl`, `foaf`, `crm`, `gvp`, and `schema` refer to well-known namespaces on the Web. `rel` contains, e.g., the schema of the application, and `nbf` is the namespace of BiographySampo. Next, the CONSTRUCT template for generating connection instances is presented in terms of variables beginning with '?'. The value bindings for the variables are determined by matching the WHERE template in all possible ways with the underlying knowledge graph. The WHERE template matches first the person and then the place and time of death. After this, a URI identifier for the connection instance is concatenated from the matched variables using the concat function of SPARQL. Finally, the natural language explanation "*?givenName ?familyName* has died in place *?placename* in the year *?year*" (in Finnish) of the connection instance is concatenated in the same way.

The form of created relation instances can be seen in the CONSTRUCT template of the above query: the class `Relation` has the following properties: type of the relation (`relationType`), the person of the relation (`personSubject`), the place of the relation (`placeObject`), the date of event (`date`), link to the underlying event (`source`), name of the underlying event source (`sourceName`), the explanation of the relation (`prefLabel`). An example of a connection instance telling that "Elin Danielson-Gambogi got the Florence City Art Award in 1899" is presented below as an example. Here the connection type is "person X received a honour related to place P".

```
a                 rel:Relation ;
rel:relationType  rel:honourAtPlace ;  # Connection type
rel:personSubject nbf:p2264 ;          # Elin Danielson-Gambogi
rel:placeObject   rel:p5133 ;          # Florence
rel:date          "1899-01-01"^^xsd:date ;  # Date of the underlying event
rel:source        nbf:event28034 ;          # Event in BiographySampo
rel:sourceName    "Tapahtuma Semanttisessa kansallisbiografiassa" ;
skos:prefLabel    "Elin Danielson-Gambogi on vastaanottanut
                   kunnianosoituksen joka liittyy paikkaan Firenze:
                   'Firenzen kaupungin taidepalkinto 1899'." .
```

Table 2 contains related connection types "Painting depicts a place" and "Novel depicts a place". These connection types can be seen to represent a more general connection "Artwork depicts a place". Instances of both of these

connection types could be created with a single SPARQL query corresponding to a more general artwork rule, but the resulting query would be more complex. We chose these connection types as a case study because these relationships were deemed interesting for the BiographySampo portal, and enough data was available in the material we had access to.

Type of Connection	**# of Connections**
Historical event in a place	345
Letter sent from	575
Letter received from	124
Text describes a place	881
Received an award in a place	2528
Died in	7349
Painting depicts a place	1091
Novel depicts a place	290
Born in	7182
Career is related to a place	20536
In total	40901

Table 2. Connection classes and their instance counts

Based on the transformed data, relational search queries can now be expressed in terms of selections on the facets and be solved efficiently using faceted search. Connection instances can now be searched for in a natural way using faceted search, where the facets are based on the property values of the instances. By making selections on the facets the result set is filtered accordingly and hit counts in the facet categories are recalculated. Facet categories were organized into hierarchies, which means that selecting a supercategory then means that all subcategories are selected with one click. For example, selecting "Finland" means that all places in Finland are automatically selected.

Data Service The data underlying BiographySampo has been published on the Linked Data Finland platform[15] [19] according to the Linked Data publishing principles and other best practices of W3C [9], including, e.g., content negotiation and provision of a SPARQL endpoint[16]. The in-use applications are based on using the endpoint.

In addition to the ready-to-use data application perspectives in the BiographySampo semantic portal, the underlying SPARQL endpoint has been applied to custom data analyses in Digital Humanities research using YASGUI[17] [27] and

[15] https://ldf.fi

[16] The homepage of the data service including, e.g., documentation of the data and pointers for linked data browsing and the SPARQL endpoint, is available at: https://www.ldf.fi/dataset/nbf

[17] https://yasgui.triply.cc

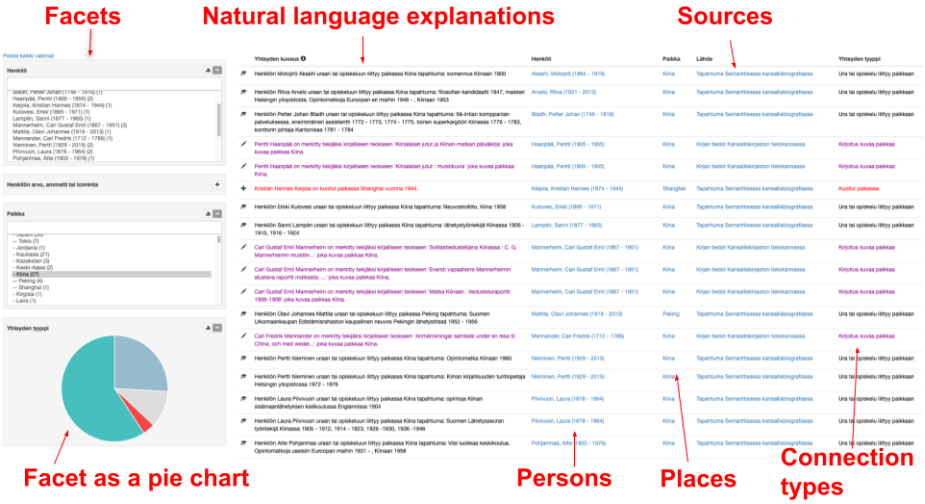

Fig. 1. View of the user interface. Facets for selections (person, occupation, place, connection type) are made on the left and results with explanations are on the right as a table rows whose columns with links correspond to the facet values.

Python scripting in Google Colab[18] and Jupyter[19] notebooks [30]. The system is compatible the "FAIR guiding principles for scientific data management and stewardship" of publishing Findable, Accessible, Interoperable, and Re-usable data are used[20].

3 Demonstrator at Work

The section shows how the application Faceted Relatoris used in practise and first evaluation results of the system.

User Interface Fig. 1 depicts the user interface of the application. The data and interface are in Finnish, but there is a Google Translate button in the right upper corner of the interface for foreign users available.

In this case study, Faceted Relator can be used for filtering relations with selections in four facets seen on the left: 1) person names, 2) occupations, 3) places, and 4) relation types. The system shows a hit list of the relation instances that fit the selected filtering criteria in the facets. The user can limit the search at any time with a selection on any facet. Furthermore, the fact that the facets are hierarchical allows searching for relations between groups of people (on the occupations facet, e.g., "film director") and larger areas (e.g., "South America") instead of individual persons or places. After each selection, the hit

[18] https://colab.research.google.com/notebooks/intro.ipynb
[19] https://jupyter.org
[20] https://www.go-fair.org/fair-principles/

counts on the facet categories tell how many results there will be it in the result set if a category is selected next. In this way, the user is guided towards filtering the solutions and never ends up in a "no hits" situation. The hit counts can also be used for visualizing the distribution of the results along each facet dimension, which is useful in quantitative analyses.

Each connection instance is represented in a row in the hit list on the right. A row shows first the natural language explanation of the connection, then the related person, place, name of the data source, and finally the connection type (cf. Table 2), based on the corresponding connection instance. Persons, places, and data sources are represented as links to further information. For example, the person link leads to the "home page" of the person in BiographySampo that automatically reassembles and visualizes the life story of the person based on the various interlinked datasets of the system. Different types of relations are highlighted in different colors and have their own symbols in order to give the user a visual overview of different kind of relations found. At any point, the distribution of the hit counts in categories along each facet can be visualized using a pie chart—one of them can be seen in the lower left corner of Fig. 1.

For example, the question "How are Finnish painters related to Italy?" is solved by selecting "Italy" from the hierarchical place facet and "painter" from the occupation facet. Any selection automatically includes its subcategories in the facet. For example, places such as Florence and Rome are in Italy, and Vatican further in Rome. The result set in this case contains 140 connections of different types whose distribution and hit counts can be seen on the connection type facet. In the same way, the person facet shows the hit count distribution along the person facet. Any facet could be used to filter the results further, if needed. In this case the 140 hits include, e.g., connection "Elin Danielson-Gambogi received in 1899 the Florence City Art Award" and "Robert Ekman created in 1844 the painting 'Landscape in Subiaco' depicting a place in Italy"[21].

In faceted search, the hit counts of facet categories tell the quantitative distributions of the results along the facet categories. This feature is utilized in Faceted Relator by making it possible to study the distributions as pie charts by clicking on a button on a facet. This feature can be used in Faceted Relator for solving some quantitative research problems.

For example, Fig. 2 illustrates how the question "Who created most painting depicting France" can be solved by selecting the connection type "Painting depicts a place" (In Finnish: "Maalaus liittyy paikkaan") on the connection type face on the bottom, and on the place facet above it "France" (In Finnish: "Ranska", including the cities, such as Paris, and other places there listed as facet subtypes). By hitting a button on the people facet, the hit distribution and pie chart along the people facet shows immediately that the female painter Ester Helenius has the most paintings of France in the available data, with 35 paintings of the total of 143 paintings that depict France. In a similar manner we could, for example, find out that general Carl Gustaf Mannerheim has most awards relating to Germany, by making the appropriate selections from the facets.

[21] These explanations are in Finnish and are translated here in English for illustration.

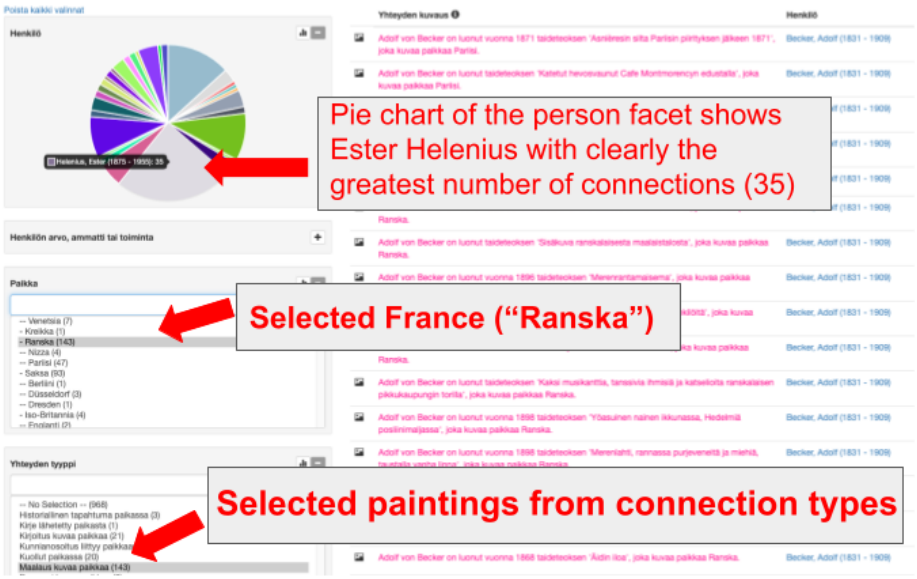

Fig. 2. Solving the problem: who created most paintings depicting France?

When using the application it is important to note that the demonstrator is limited by the sources and data it uses. A relation can be missing for a number of reasons and relative numbers may not therefore reflect reality perfectly. However, the tool can be valuable for finding out serendipitous phenomena in the data for further close reading by the human expert.

Evaluating the System To evaluate the quality of the relations and explanations given by the system we evaluated the results received with a small number of searches. We made a search to find relations starting from five different people and places.

The people selected were: (1) Elias Lönnrot (1802–1884), the creator of the Finnish national epic Kalevala, (2) Johan Ludvig Runeberg (1804–1877), the Finnish national poet, (3) Akseli Gallen-Kallela (1865–1931), one of the most prominent Finnish classical painters, (4) Ellen Thesleff (1869–1954), a female Finnish expressionist, and (5) Urho Kekkonen (1900–1986), the longest serving president of the Finland. After selecting a person in the person facet, Faceted Relator determined the related connections that were analysed manually.

These people were selected for their significance to the Finnish history representing different fields and times. They do not represent the average people in the data but were expected to have many relations of diiferent kinds to places for evaluation in the data. The searches with the selected people yielded from 18 to 44 relations to places for each.

1. Elias Lönnrot has 42 relations, including, some places related to his career as a doctor of medicine in Oulu and Kajaani. He has received the Prussian

Pour le Merite award. There are many letters, including some that he has sent from Estonia, relating him to places where he wrote or received letters. He is also the author of a few books concerning certain Finnish places. All the relations found seem to represent Lönnrot's life quite well, and the natural language explanations were good, too. A few of his books are mentioned multiple times, because they were published in multiple languages.

2. Johan Ludvig Runeberg has 18 relations of multiple types. For example he has received Danish and Swedish honorary medals. Most of the explanation seem good, but one relation concerning his career is perhaps misleading. According to the system, Runeberg's career is related to Greece, because he was a teacher of classical Greek language at one point. This isn't entirely wrong, but might be seen as misleading.
3. Akseli Gallen-Kallela has 44 relations. These concern only his birth, death and, his paintings, and certain books about his painting. Interestingly twenty of these are related to Africa, far more than, for example, to Paris. This may overstate the meaning of Africa for Gallen-Kallela's life, but it does reflect the fact that did spend almost two years in Africa. This can also be surprising information to someone with only passing information about Gallen-Kallela and inspire the user to learn more about him. It is notable that the system doesn't show any career events for Gallen-Kallela. This seems to be because the biography of Gallen-Kallela is structured in such a way that no career events were picked to the Biographysampo knowledge graph, and therefore no relations can be generated based on them.
4. Ellen Thesleff has 40 relations. These concern her birth, death, and paintings. Many of the painting are related to Italy, which does reflect the importance of Italy for her work. It is notable that both Thesleff and Gallen-Kallela lack relations concerning their career and awards in the system. Both of their biographies certainly include many interesting career events and awards that could, and ideally should, be included. These are lacking because the these events are collected to the Biographysampo knowledge graph from certain sections of the biographies, that are lacking with Gallen-Kallela and Thesleff.
5. Urho Kekkonen has 29 relations. These include some historical events, incuding his presidential election, and notably many honours he has received in the form of honorary doctorates from around Finland and the World. Most of these have good natural language explanations, but few have somewhat mysterious looking explanations like "Urho Kekkonen received an honor related to Varsova: Varsova 1964". This refects the fact that Kekkonen received an honorary doctorate from Varsova in 1964, and therefore it is not wrong but the explanation is not good. This happens because in the biography certain types of honorary doctorates are given as a list. It might be possible to eliminate these kind of vague explanations when creating the relation entities, for example by automatically excluding all awards with too short explanations. However then these potentially interesting connections would not be shown. Notably the birth place of Kekkonen is missing. This is likely due to an omission in the mapping of place ontologies.

We also searched connections by starting from five places. The selected places are (1) Utsjoki, the most northernmost town of Finland to represent a smaller place, (2) Helsinki and (3) Turku, the two most important cities of the Finnish history, and (4) London and (5) Paris, representing important cities outside Finland. After selecting a place in the place facet, Faceted Relator determined the related connections that were analysed manually. Searching for people related to a certain place the user should first select a place from the place facet.

1. Utsjoki has only 8 relations, so there is no need for narrowing the search as all the explanations can be easily read. There is a variety of different facts and this could well be used to find out about the local history of the town.
2. There are more than 8000 connections for Helsinki, more than to any other place, which can be expected for the capital of Finland. When the user selects Helsinki, he is shown all the connections as a list ordered by name of the person in the connection. The number of connections is too large to go through and read them all. The lack of prioritization means that the user may not find interesting connections by just looking at the results. Here the user needs to explore the facets and narrow the search further to find interesting individual connections. Here the system is working as planned and invites the user to explore the data interactively, but some users might want a ranked selection of connections, so that they would be immediately offered most interesting results on top of the search results. To limit the results the user could further narrow the search to, for example, people of certain profession. However even without narrowing the search further, the user could compare the relative numbers of relations using the pie chart option and see that Helsinki has a relatively large number of connections to members of the Parliament and authors.
3. Turku has over 3000 connections, and there is a need to narrow the search further as in the case of Helsinki. An interesting result are the relative numbers of the connections on the facets that can be visualized with a pie chart. Especially interesting might be a comparison with Helsinki. For example, Turku has a relatively larger number of connections to priests. The pie chart shows that the profession to which both Helsinki and Turku have most connections is Member of the Parliament.[22]
4. London has 171 connections, and this might also be a too large number to go through and might require further narrowing the search to. For example, to find out how authors are related to London, an additional selection on the occupation/profession facet is needed. This would reveal, among other things, that the author Aale Tynni won a gold medal in poetry in the London Olympic Games of 1948. This is an example of serendipitous piece of information to those who do not know that poetry used be a competition in the Olympic Games.
5. Paris has 446 connections in the system. Again there are a lot of connections, but the relative numbers might be interesting even without further narrowing

[22] A single person may have several connections to a place summed up here.

of the search. These can be compared to other places such as London. The user could, for example, compare the profession distribution, and find out that Paris has more connections to painters than London. Also the fact that Paris has more connections altogether can be interesting. It hints that Paris has been more culturally significant for Finland than London.

The informal evaluation and testing above, as well as some additional tests, showed that the method and system works as well in terms of precision. This was not a big surprise, as the connections in our method are determined by explicit logical rules. As for recall, evaluation of the results is challenging, as there is no golden standard available, and failing to find a connection may be due to sparsity of the data, not the method. In any case, as the Table 2 shows, the system was able to find lots interesting relations in the data and the approach looks promising. Theoretically it seems likely that this kind of approach will miss some truly serendipitous connections that represent some type of relation that could not be even thought of. This is because the nature of the method requires limiting the search to predetermined types of connections. It could be argued that this method gives preference to precision over recall.

According to [4], a system can be considered creative if it is able to create "new", "surpising", and "valuable" ideas. At least from a layman perspective, this seems to be the case in Faceted Relator although measuring creativity is not easy. Given the large, semantically rich knowledge graph we believe that the system can provide insightful results even for an expert historian. However, more testing is needed to find out how interesting and surprising the results are for an expert of CH and how a system like this can be used for DH research.

Generalizability of the Knowledge-based Approach In our example, we have searched for connections between prominent Finnish people and places. Generalizing the method to other application domains and datasets would in principle be straightforward, but in practice adaptation work is needed as the data about people and places in other datasets may be different, and it may also be represented using different data models and ontologies. If the data in several datasets is represented using standard data and vocabulary models, such as CIDOC CRM and SKOS, the same rules for instantiating connections can be re-used in different datasets.

More work would be needed to apply the method to different relations, such as relations between two people, pieces of art, or events. Different types of relations would require domain and dataset specific considerations. For example, when generating interesting connections between two people a connection "two prominent people were born in the same city" may generate a huge number of relations that would be both difficult to search efficiently and not generally particularly interesting if most people are born in the same few large cities. More interesting relation might be, for example, "two prominent people were born in the same small town or village". This would keep the number of connections relatively low and the individual connections are more likely to be interesting. A weakness of knowledge-based methods like ours is the need to customize the method to fit to specific case and data—on the other hand fitting the method

to particular applications is also a strength of the approach as non-interesting connections can be ruled out.

The number of connection instances needs to be limited to allow for efficient faceted search in interactive usage with a few seconds response times. A most computationally demanding part is counting of instances for all possible facet selections after each selection. According to our tests with Faceted Relator some 40 000 instances could be searched efficiently by faceted search using a database server corresponding to a normal personal computer. The needed efficiency is however dependent on the application. In our case, the application is mostly aimed for the general public and needs to work relatively fast. A system aimed for professional audience might allow for longer search times. The queries used in creating the connection instances are run in a separate preprocessing phase and do not make response times longer when using the portal but wise versa improves real time efficiency.

4 Discussion

Creating connections and explanations for entities based on knowledge graphs is relatively simple if the data is semantically rich enough. Making general forms that would fit to various data sources is not currently feasible, because the standards are not generally used in uniform manner, but creating new forms for each new data model is usually not challenging. Hierarchical ontologies relating to the entities makes it easy to then search for connections for larger groups, such as groups of people (e.g., painters) and larger geographical areas (e.g., Italy). It was surprising to us how useful this approach of finding connections for larger groups using faceted search was. Currently the natural language explanations for connections are only for single entities, but it would be an interesting and challenging research question how to create summarized natural language explanations for the connections between groups of entities.

The endpoints of a connection can be seen as equivalent to subject and object in an RDF triple, while the type of the connection can be seen as property in a triple. The connections can be grouped and searched based on the properties of each of these elements. In our demonstrator we have only a limited number if facets, but they demonstrate filtering for groups based on each of the three elements, with the occupation facet, hierarchical place facet, and the facet for the connection type. It would be possible to have more facets for each of these, but that would obviously require more ontological infrastructure.

The knowledge acquisition task of formulating a set of useful explanation patters and graph transformation rules in the demonstrator was feasible. Furthermore, the number of connections found was not overwhelmingly large from a computational point of view, and could be generated quickly. From a human end user perspective, the result set (40 901 connections) is still large enough to provided many non-trivial results and explanations. So, the suggested knowledge-based approach was deemed feasible at least in cases where the potentially in-

teresting connections can be characterized logically and their number is not very large. This seems to be the case in the biographical datasets of BiographySampo.

Faceted search can be used to narrow the search for those relations that the user would consider most interesting. The incremental nature of faceted search makes it more likely to make serendipious discoveries, but it might be useful to augment the search with some sort of ranking of relations based on their presumed interestingness. This would be especially important in cases where the number of relation instances is very high.

If the constraints on interestingness, i.e., the transformation rules, are loosened too much, there is the danger for combinatorial explosion of results, and very common connections would probably not be very interesting. This should therefore be avoided. For example, the connection that two person are born in the same country would connect most the people in our data, and would not be interesting and worth generating. However, if the persons are born in a small village and at about the same time, the connection would be much more interesting. Using a more refined connection type can reduce the number of connections. For example, one could search for connected persons born in a small village or born around the same time in a larger community. We believe that domain knowledge is useful and in many cases necessary in making such fine-grained distinctions of interestingness.

Future Work When testing and evaluating the demonstrator, we also found out needs to improve the usability of the system. For example, the demonstrator now sorts results based on firstly the name of the person and secondly on the name of the place. The user should probably be offered the possibility to sort the relations freely along any facet. Developing the ontologies, such as the ontology of professions, might also improve usability of the system. We would like to expand the system with new connection types, such as relations between people. We are also planning on expanding the approach to data that covers people from other European countries.

Acknowledgements Our research was originally supported by the Severi project[23], funded mainly by Business Finland, and then by the EU project InTaVia: In/Tangible European Heritage[24]. The work was also part of the *Open Science and Research Programme*[25], funded by the Ministry of Education and Culture of Finland. The authors wish to acknowledge CSC – IT Center for Science, Finland, for 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

[23] http://seco.cs.aalto.fi/projects/severi
[24] https://intavia.eu/
[25] http://openscience.fi

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. Birró, G.: Building relatedness explanations from knowledge graphs. Semantic Web **10**(6), 963–990 (2020)
4. Boden, M.A.: Computer models of creativity. AI Magazine **30**(3), 23–34 (2009), http://www.aaai.org/ojs/index.php/aimagazine/article/view/2254
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. Došilović, F.K., Brčić, M., Hlupić, N.: Explainable artificial intelligence: A survey. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). pp. 210–215. Rijeka, Croatia (2018)
8. Gutierrez, C., Sequeda, J.F.: Knowledge graphs. Communications of the ACM **64**(3), 96–104 (March 2021). https://doi.org/10.1145/3418294
9. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Morgan & Claypool, Palo Alto, California (2011), http://linkeddatabook.com/editions/1.0/
10. 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
11. 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
12. Herlocker, J.H., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Computer Supported Cooperative Work. pp. 241–250. ACM (2000)
13. Hyvönen, E.: Publishing and using cultural heritage linked data on the Semantic Web. Morgan & Claypool, Palo Alto, California (2012). https://doi.org/10.2200/S00452ED1V01Y201210WBE003
14. 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
15. 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
16. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., Keravuori, K.: BiographySampo - publishing and enriching biographies on the se-

mantic web for digital humanities research. In: Proceedings of the 16th Extendwed Semantic Web Conference (ESWC 2019). Springer–Verlag (2019)
17. Hyvönen, E., Rantala, H.: Relational search in cultural heritage linked data: A knowledge-based approach. In: Digital Humanities 2019 Conference Papers, Book of Abstracts. University of Utrecht (2019), https://dev.clariah.nl/files/dh2019/boa/0445.html
18. 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
19. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: The Semantic Web: ESWC 2014 Satellite Events. pp. 226–230. Springer–Verlag (May 2014). https://doi.org/10.1007/978-3-319-11955-7_24
20. Khalili, A., van Andel, P., van den Besselaar, P., de Graaf, K.A.: Fostering serendipitous knowledge discovery using an adaptive multigraph-based faceted browser. In: Proceedings of the Knowledge Capture Conference. K-CAP 2017, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3148011.3148037, https://doi.org/10.1145/3148011.3148037
21. 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
22. Lecue, F.: On the role of knowledge graphs in Explainable AI. Semantic Web – Interoperability, Usability, Applicability **11**(1), 41–51 (2020)
23. 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
24. 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
25. Maimon, O., Rokach, L. (eds.): The data mining and knowledge discovery handbook. Springer–Verlag (2005)
26. Mäkelä, E., Ruotsalo, T., Hyvönen: How to deal with massively heterogeneous cultural heritage data—lessons learned in CultureSampo. Semantic Web – Interoperability, Usability, Applicability **3**(1), 85–109 (2012)
27. Rietveld, L., Hoekstra, R.: The YASGUI family of SPARQL clients. Semantic Web – Interoperability, Usability, Applicability **8**(3), 373–383 (2017). https://doi.org/10.3233/SW-150197
28. 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)
29. Silberschatz, A., Tuzhilin, A.: On subjective measures on interestingness in knowledge discovery. In: Proceedings of KDD-1995. pp. 275–281. AAAI Press (1995)
30. 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 (2021), https://seco.cs.aalto.fi/publications/2021/tamper-et-al-bs-2021.pdf, forth-coming

31. 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)
32. 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 Workshop Proceedings (2018), http://ceur-ws.org/Vol-2119/paper10.pdf