Finding and explaining relations in a biographical knowledge graph based on life events: Case BiographySampo

Heikki Rantala[1], Eero Hyvönen[1,2] and Petri Leskinen[2,1]

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

Abstract
This paper presents a knowledge-based approach for finding and explaining "interesting" semantic relations between persons and places in a knowledge graph. As a case study we use the BiographySampo knowledge graph which includes life events extracted from the short biographies of 13 100 prominent historical persons in Finland. We use SPARQL CONSTRUCT queries to extract connections and create human readable explanations based on the events in the knowledge graph, and then offer faceted search tools to search and visualize the connections based on various criteria.

Keywords
knowledge discovery, linked data, event-based model

1. Knowledge Discovery as Relational Search

Research Problems This paper addresses the following problem of knowledge discovery [1] in Cultural Heritage (CH) [2] knowledge graphs (KG) [3]: *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 as association finding [4] or as relational search (RS) [5, 6, 7, 8]. We address the following challenges of solving RS problems:

1. *How to disambiguate "interesting" [9] or even "serendipitous"[1] [10] 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.

heikki.rantala@aalto.fi (H. Rantala); eero.hyvonen@aalto.fi (E. Hyvönen); petri.leskinen@aalto.fi (P. Leskinen)
https://seco.cs.aalto.fi/u/rantalh3/ (H. Rantala); https://seco.cs.aalto.fi/u/eahyvone/ (E. Hyvönen); https://seco.cs.aalto.fi/u/ptleskin/ (P. Leskinen)
0000-0002-4716-6564 (H. Rantala); 0000-0003-1695-5840 (E. Hyvönen); 0000-0003-2327-6942 (P. Leskinen)

[1]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".

2. *How to explain a semantic connection to the end user?* Finding out an interesting connection is not enough if the system cannot explain to the end user why the connection could be interesting. This problem is related to the field of explainable AI [11, 12].
   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[13], 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.*

Related Works In relational search the *query* consists of two or more resources, and the task is to find interesting semantic relations between them. The approaches [14] differ in terms of the query formulation, underlying KG, methods for finding connections, and representation of the results. In [4] the idea of searching relations is applied for association finding in national security domain. CultureSampo[2] [15, 16] contains an application where connections between two persons were searched using a breadth-first algorithm. In RelFinder[3] [17, 5, 6, 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 [8] finds several paths with a relevance measure between two resources in the WikiData KG[4], 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 [18]. Some applications, e.g., RelFinder and Explass [19], allow filtering relations between two entities with facets, but the user typically has preselect the entities before faceted search can be used. A main challenge in these systems is how to select and rank the interesting paths. Ranking relations is discussed, e.g., in [14, 20].

In [21] 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 [22].

This paper presents and applies a knowledge-based approach to the research problems above and presents experiences in applying the approach in BiographySampo[23, 24]"BiographySampo – Finnish Biographies on the Semantic Web"[5], which includes biographical data about historical Finnish persons expressed as Linked Open Data using event based Bio CRM [25] model, an extension of CIDOC CRM[6] designed for biographical data. This paper extends and complements our earlier papers [26, 27].

[2] http://www.kulttuurisampo.fi
[3] http://www.visualdataweb.org/relfinder.php
[4] http://wikidata.org
[5] Project: https://seco.cs.aalto.fi/projects/biografiasampo/; portal: https://biografiasampo.fi/, online since 2018
[6] http://cidoc-crm.org

The biographical data of BiographySampo is based on 13 144 biographies and includes thousands of life events, including births, deaths, career events, received accolades, and even historical events where the persons have participated in. The life of each biographee was 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. [24] BiographySampo has also been enriched from other sources, such as the HISTO[7] ontology of Finnish historical events. These event can create various types of connections between persons and other persons or places. For example two people might have been born in the same place or participated in the same historical event.

2. Finding and Explaining Semantic Relations

While the graph-based methods above make use of generic graph traversal algorithms that are application domain agnostic, our method uses 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 [27] re-using a ready-to-use tools [28, 29] 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. However, in this method the transformation rules and their explanation patterns need to be crafted manually, based on application domain knowledge.

Below is an example of a CONSTRUCT QUERY used to create Relation instances. The query[8] finds pairs of two people who have both participated in the same historical event, 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. An example of an Finnish language explanation generated is “Rentola, Rauha ja Larkka, Erkki osallistuivat samaan historialliseen tapahtumaan: Suomen ensimmäinen julkinen televisiolähetys”, which can be translated as “Rentola, Rauha and Larkka, Erkki took part in the same historical event: the first public television broadcast in Finland”. There are 1934 distinct Relation instances created by this query. It is good to note that because the connections are directed, essentially the same connections are extracted twice so that both persons are the starting point of the connection once. The queries are not too computationally demanding, and executing queries like the one below usually only takes a couple of seconds. The example given here is a minimal one. The Relation instances can also include semantic information about, for example, times and sources of the connections.

```
PREFIX bioc:    <http://ldf.fi/schema/bioc/>
PREFIX foaf:    <http://xmlns.com/foaf/0.1/>
PREFIX rdfs:    <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skosxl:  <http://www.w3.org/2008/05/skos-xl#>
```

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

[8] You can test the query in Yasgui https://api.triplydb.com/s/z_9sI7Fox.

```
PREFIX skos:    <http://www.w3.org/2004/02/skos/core#>
PREFIX nbf:     <http://ldf.fi/nbf/>
PREFIX rel:     <http://ldf.fi/schema/relations/>
CONSTRUCT {
  [] a rel:Relation ;
    rel:relationSubject ?person_A ;
    rel:relationObject ?person_B ;
    rdfs:label ?description ;
    rel:relationType rel:sharedEvent .
  }
WHERE {
  ?event a nbf:Event .
  ?event skos:prefLabel ?event_label .
  ?event bioc:inheres_in ?person_A.
  ?event bioc:inheres_in ?person_B.
  FILTER (?person_A != ?person_B)
  ?person_concept_A foaf:focus ?person_A .
  ?person_concept_B foaf:focus ?person_B .
  ?person_concept_A skosxl:prefLabel/skos:prefLabel ?A_label .
  ?person_concept_B skosxl:prefLabel/skos:prefLabel ?B_label
  BIND(CONCAT(CONCAT(CONCAT(CONCAT(?A_label, ' ja '),
    ?B_label), ' osallistuivat samaan historialliseen tapahtumaan: '),
    ?event_label) AS ?description)
}
```

3. Faceted Search User Interface

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[9]. This application is based on SPARQL Faceter[28], and is partially documented in [27]. We are now working on a more general case where an application could be used to find connections between two people or groups of people. We are initially working on only Finnish persons from the BiographySampo KG, but we are planning to include international data as well from other European countries. The new web application will be based on Sampo-UI[29], and we plan to publish it later in 2023.

In our 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. Figure 1 shows an example of the user interface. The facets are located on the left side of the scree and the human readable explanations of each relation are shown on the right as well as relevant links to the entities of the relation. A user can simply select a single entity, a place or a person, from a facet, and then look at the various relations that entity has to others. User can however also search for larger groups by, for example, making a selection from “Occupation” facet, so that the user is shown all the relation where the person has certain occupation. Similarly, a place facet can be hierarchical, so that the user can search for larger area than the place that is directly part of the connection. For example user can select Italy, and get connections for Rome and other places in Italy.

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 by making it possible to study the distributions as pie charts by clicking on a button on a facet. This feature can be used for

[9] http://biografiasampo.fi/yhteyshaku/

solving some quantitative research problems. For example, Fig. 1 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 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.

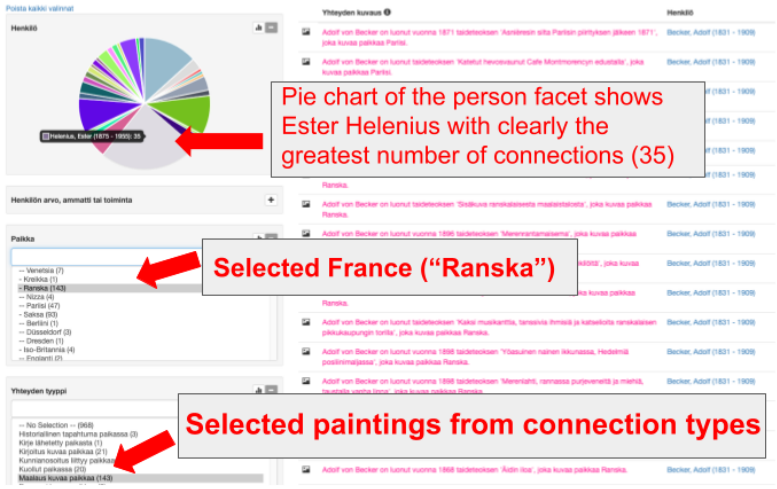

Figure 1: Example of using the user interface. Solving the problem: who created most paintings depicting France?

4. Discussion and future work

While finding single connections and their explanations between two entities can be interesting, often more interesting are the larger the larger connections, that can be seen through faceted search and visualizations. The larger 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 first demonstrator, because that is a simpler case. Working on connections between persons offers new challenges. These challenges are related to both defining the relation types so that the number of Relations stays manageable for faceted search, which can be resource consuming if the number of searched entities is large, and to implementing the faceted search user interface.

When searching connections between different types of entities like people and places, it is easy for user the 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 faceted 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 double the number of Relation instances which can slow the faceted search.

We are working on creating a demonstrator later in 2023 for searching connections between persons which is first applied to BiographySampo, and then extended to use international data from biographies from other European countries. We are also working on apllying the approach to other cases, such as the artist data of the GETTY ULAN [10] KG.

Acknowledgments Our research was partly supported by the EU project InTaVia[11]. CSC – IT Center for Science, Finland, provided computational resources.

References

[1] O. Maimon, L. Rokach (Eds.), The data mining and knowledge discovery handbook, Springer–Verlag, 2005.

[2] E. Hyvönen, Publishing and using cultural heritage linked data on the Semantic Web, Morgan & Claypool, Palo Alto, California, 2012. doi:10.2200/S00452ED1V01Y201210WBE003.

[3] C. Gutierrez, J. F. Sequeda, Knowledge graphs, Communications of the ACM 64 (2021) 96–104. doi:10.1145/3418294.

[4] A. Sheth, B. Aleman-Meza, I. B. Arpinar, C. Bertram, Y. Warke, C. Ramakrishnan, C. Halaschek, K. Anyanwu, D. Avant, F. S. Arpinar, K. Kochut, Semantic association identification and knowledge discovery for national security applications, Journal of Database Management on Database Technology 16 (2005) 33–53.

[5] S. Lohmann, P. Heim, T. Stegemann, J. Ziegler, The RelFinder user interface: Interactive exploration of relationships between objects of interest, in: Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI 2010), ACM, 2010, pp. 421–422. URL: http://doi.acm.org/10.1145/1719970.1720052.

[6] P. Heim, S. Lohmann, T. Stegemann, Interactive relationship discovery via the semantic web, in: Proceedings of the 7th Extended Semantic Web Conference (ESWC 2010), volume 6088, Springer–Verlag, Berlin/Heidelberg, 2010, pp. 303–317. URL: http://dx.doi.org/10.1007/978-3-642-13486-9_21.

[7] P. Heim, S. Hellmann, J. Lehmann, S. Lohmann, T. Stegemann, Relfinder: Revealing relationships in rdf knowledge bases, in: Proceedings of the 4th International Conference on Semantic and Digital Media Technologies (SAMT 2009), Springer–Verlag, 2009, pp. 182–187. URL: http://dx.doi.org/10.1007/978-3-642-10543-2_21.

[8] G. Tartari, A. Hogan, WiSP: Weighted shortest paths for RDF graphs, in: Proceedings of VOILA 2018, CEUR Workshop Proceedings, vol. 2187, 2018, pp. 37–52.

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

[11] https://intavia.eu/

[9] A. Silberschatz, A. Tuzhilin, On subjective measures on interestingness in knowledge discovery, in: Proceedings of KDD-1995, AAAI Press, 1995, pp. 275–281.
[10] A. Khalili, P. van Andel, P. van den Besselaar, K. A. de Graaf, 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. URL: https://doi.org/10.1145/3148011.3148037. doi:10.1145/3148011.3148037.
[11] F. K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: A survey, in: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Rijeka, Croatia, 2018, pp. 210–215.
[12] F. Lecue, On the role of knowledge graphs in Explainable AI, Semantic Web – Interoperability, Usability, Applicability 11 (2020) 41–51.
[13] D. Tunkelang, Faceted search, Synthesis Lectures on Information Concepts, Retrieval, and Services 1 (2009) 1–80.
[14] G. Cheng, F. Shao, Y. Qu, An empirical evaluation of techniques for ranking semantic associations, IEEE Transactions on Knowledge and Data Engineering 29 (2017) 1.
[15] E. Hyvönen, E. Mäkelä, T. Kauppinen, O. Alm, J. Kurki, T. Ruotsalo, K. Seppälä, J. Takala, K. Puputti, H. Kuittinen, K. Viljanen, J. Tuominen, T. Palonen, M. Frosterus, R. Sinkkilä, P. Paakkarinen, J. Laitio, K. Nyberg, 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. URL: https://seco.cs.aalto.fi/publications/2009/hyvonen-et-al-culsa-mw-2009.pdf.
[16] E. Mäkelä, T. Ruotsalo, Hyvönen, How to deal with massively heterogeneous cultural heritage data—lessons learned in CultureSampo, Semantic Web – Interoperability, Usability, Applicability 3 (2012) 85–109.
[17] J. Lehmann, J. Schüppel, S. Auer, Discovering unknown connections—the DBpedia relationship finder, in: Proc. of the 1st Conference on Social Semantic Web (CSSW 2007), volume 113 of *LNI*, GI, 2007, pp. 99–110. URL: http://subs.emis.de/LNI/Proceedings/Proceedings113/gi-proc-113-010.pdf.
[18] M. Al-Tawil, V. Dimitrova, D. Thakker, Using knowledge anchors to facilitate user exploration of data graphs, Semantic Web 11 (2020) 205–234. doi:10.3233/SW-190347.
[19] G. Cheng, Y. Zhang, Y. Qu, Explass: exploring associations between entities via top-k ontological patterns and facets, in: International Semantic Web Conference (ISWC), Springer–Verlag, 2014, pp. 422–437.
[20] F. Bianchi, M. Palmonari, M. Cremaschi, E. Fersini, Actively learning to rank semantic associations for personalized contextual exploration of knowledge graphs, in: E. Blomqvist, D. Maynard, A. Gangemi, R. Hoekstra, P. Hitzler, O. Hartig (Eds.), The Semantic Web, Springer–Verlag, Cham, 2017, pp. 120–135. doi:10.1007/978-3-319-58068-5_8.
[21] G. Birró, Building relatedness explanations from knowledge graphs, Semantic Web 10 (2020) 963–990.
[22] J. H. Herlocker, J. A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Computer Supported Cooperative Work, ACM, 2000, pp. 241–250.
[23] E. Hyvönen, P. Leskinen, M. Tamper, H. Rantala, E. Ikkala, J. Tuominen, K. Keravuori, 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.

[24] M. Tamper, P. Leskinen, E. Hyvönen, R. Valjus, K. Keravuori, Analyzing biography collection historiographically as linked data: Case national biography of finland, Semantic Web – Interoperability, Usability, Applicability 14 (2023) 385–419. URL: https://doi.org/10.3233/SW-222887.

[25] J. Tuominen, E. Hyvönen, P. Leskinen, Bio CRM: A data model for representing biographical data for prosopographical research, in: BD2017 Biographical Data in a Digital World 2017, Proceedings, CEUR WS Proceedings, 2018, pp. 59–66. http://ceur-ws.org/Vol-2119/paper10.pdf.

[26] E. Hyvönen, H. Rantala, Relational search in cultural heritage linked data: A knowledge-based approach, in: Digital Humanities 2019 Conference Papers, Book of Abstracts, University of Utrecht, 2019. URL: https://dev.clariah.nl/files/dh2019/boa/0445.html.

[27] E. Hyvönen, H. Rantala, Knowledge-based relational search in cultural heritage linked data, Digital Scholarship in the Humanities (DSH) 36 (2021) 155–164. doi:https://doi.org/10.1093/llc/fqab042.

[28] M. Koho, E. Heino, E. Hyvönen, SPARQL Faceter – Client-side faceted search based on SPARQL, in: Joint Proceedings of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop, CEUR Workshop Proceedings, 2016, pp. 53–63. URL: http://ceur-ws.org/Vol-2187/paper5.pdf.

[29] E. Ikkala, E. Hyvönen, H. Rantala, M. Koho, Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces, Semantic Web – Interoperability, Usability, Applicability 13 (2022) 69–84.