Aggregating and Aligning Knowledge Graphs into a Global Service: SampoSampo System for Cross-cultural Data Search, Exploration, and Analysis*

Eero Hyvönen[1,2,*], Annastiina Ahola[1], Petri Leskinen[1] and Jouni Tuominen[2,3]

[1]Aalto University, Department of Computer Science
[2]University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG)
[3]University of Helsinki, Helsinki Institute for Humanities and Social Sciences (HSSH)

Abstract

This paper abstract presents an approach and first results of creating a global LOD service and semantic portal, SampoSampo, based on a network of interlinked Cultural Heritage knowledge graphs (KG) of different application domains. In this way, a more comprehensive global view for searching, exploring, and analyzing the interlinked KGs can be provided than by using local KGs separately. The SampoSampo LOD service can be used as a web service for providing IRI identifiers for aligning new datasets with the national DARIAH-FI research infrastructure on which SampoSampo is based, and for assessing the quality of data in different KGs by comparing their metadata. The portal underway can be used for searching and exploring the local KGs by with a single user interface (UI) and for implementing novel application perspectives based on relational search, where semantic "interesting" associations (relations) between entities, such as persons, organizations, and places in the global KG, can be searched for and natural explanations for them be created.

Keywords

linked data, biographical data, epistolary data, semantic portal, data analysis,

1. Enriching and harmonizing data by linking

One of the great promises of Linked Data, as promoted by the 5th star in Tim Berners-Lee 5-star model[1], is enriching data by linking it to external datasets using IRIs. A requirement for this is that the linked datasets use the same identifiers (IRIs) for the same resources or that IRI alignments across datasets are available. For data models and Knowledge Organization System (KOS) schemas, such as Dublin Core[2], RDF[3], SKOS[4], etc., harmonized use of IRIs is often the

Digital Humanities in Nordic and Baltic Countries, Tartu, Estonia, March 5–7, 2025
* Abstract proposal for a long paper.
*Corresponding author.
eero.hyvonen@aalto.fi (E. Hyvönen)
https://seco.cs.aalto.fi/u/eahyvone (E. Hyvönen)
0000-0003-1695-84 (E. Hyvönen)
© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441
[1]5-star model: https://5-star.info
[2]Dublin Core metadata initiative: https://dublincore.org
[3]Resource Description Framework RDF: https://www.w3.org/RDF/
[4]Simple Knowledge Organization System SKOS: https://www.w3.org/2004/02/skos/

case due to standardization efforts, but not so often for data resources representing things of the real world, such as persons, organizations, and places.

This key problem of aligning and interlinking datasets based of different identifier systems has been addressed before in many ways. For example, in the library world, different identifiers are often used for, e.g., authors and places in national collections, and linking systems, such as the Virtual International Authority File service VIAF[5] have been created to mitigate the problem (Hickey and Toves 2014). Linked Open Vocabularies (LOV) is an example of a high-quality catalogue of reusable vocabularies, their alignment, and version histories (Dumontier et al. 2017).

This paper presents work on creating a new kind of VIAF-like mapping system and semantic portal called SampoSampo for cross-domain CH collections. The novelty in our case is two-fold: The system is based on existing KGs and LOD services already available on the Semantic Web, and includes not only a LOD service but a semantic portal, too. The focus is on using data of the over 20 Sampo KGs (Hyvönen 2022) that publish LOD in different application domains of Cultural Heritage, such as artifacts (CultureSampo), literature (BookSampo), musical performances (OperaSampo), artworks (ArtSampo), parliamentary speeches and networks (ParliamentSampo), culturally significant people (BiographySampo), academic people registers (AcademySampo), military history (WarSampo), and epistolary collections (LetterSampo). These KGs constitute a component of the Finnish DARIAH-FI research infrastructure[6] (Hyvönen 2024). We present the underlying ideas of SampoSampo, methods used, and report first results of creating the SampoSampo in practise.

2. Use cases of SampoSampo

There are many reasons and use cases for creating SampoSampo. The entity resources used in the Sampo systems are often based on the same infrastructural resources available at the Finnish ontology services ONKI[7] and Finto[8]. However, due to historical and other reasons, also application specific KOS have been used for populating the metadata models, and the data is linked to international datasets, too. A goal of SampoSampo is to create a kind of universal reference service and a SPARQL endpoint for using the resources of Sampos. This kind of LOD service is useful when creating and aligning new datasets with existing ones by FAIR principles[9]. Furthermore, one benefit of collecting and assembling biographical data from multiple distinct sources is getting it enriched. As an example, one data source might have the main focus on someone's career as a politician or as an artist while there might be more information about, e.g., her or his family relations in another data source. Besides that, comparing actor data imported from multiple databases also facilitates to detecting contradictions and errors in the data sources.

We also aim at providing the end users with a single semantic portal and UI to a set of underlying KGs based on their shared resources. Using this portal, it is possible to search, browse, and analyse several local KGs on a global level at the same time. Furthermore, the

[5]Virtual International Authority File system: https://viaf.org
[6]LOD part of the FIN-CLARIAH/DARIAH infrastructure: https://seco.cs.aalto.fi/projects/fin-clariah/
[7]ONKI service: https://onki.fi
[8]Finto service: https://finto.fi
[9]FAIR principles: https://www.go-fair.org/

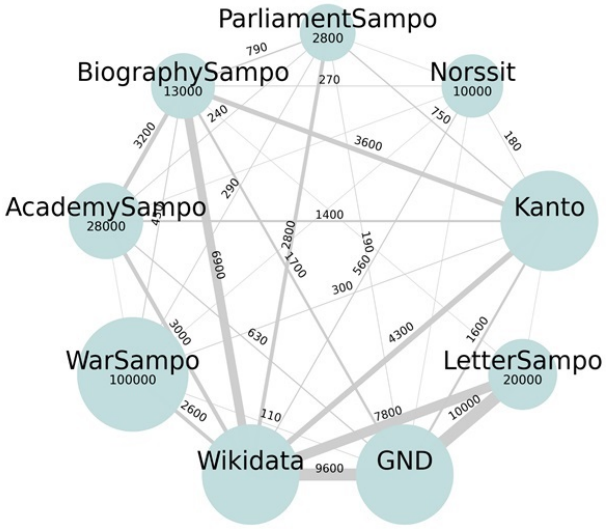

Figure 1: Interlinked actor resources in some biographical Sampos and beyond (Leskinen 2024)

portal will be capable for cross-cultural knowledge discovery of semantic association between entities (Lehmann, Schüppel, and Auer 2007; Tartari and Hogan 2018) and explaining them in natural language, using the "knowledge-based" approach to relational search (Hyvönen and Rantala 2021; Rantala, Hyvönen, and Leskinen 2023; Rantala et al. 2024)

Our focus is on resources for historical persons, organizations, and places that are widely used in virtually all Sampo systems. For example, Figure 1 illustrates the linkedness of some biographical Sampo systems and other systems, including the Kanto authority file system[10] by the National Library of Finland, Wikidata[11], and the German Integrated Authority File system of the Deutsche National Bibliothek (GND)[12]. The numbers on the connection arcs tell the number of shared resources between the connected datasets. For example, from the 100 000 persons in the WarSampo KG (Koho et al. 2021; Hyvönen et al. 2016) 2600 can be found in Wikidata and 290 in the ParliamentSampo KG (Leskinen, Hyvönen, and Tuominen 2021; Hyvönen et al. 2024).

Acknowledgements

This research was funded by the Research Council of Finland and is part of the FIN-CLARIAH initiative that has received funding from the European Union NextGenerationEU instrument. Computing resources provided by the CSC – IT Center for Science were used in our work.

[10]Kanto authorities: https://finto.fi/finaf/en/
[11]WIkidata: https://wikidata.org
[12]Gemainsame Normdatei : https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html

References

Dumontier, Michel, Pierre-Yves Vandenbussche, Ghislain A. Atemezing, María Poveda-Villalón, and Bernard Vatant. 2017. "Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web." *Semant. Web* (NLD) 8, no. 3 (January): 437–452. ISSN: 1570-0844. https://doi.org/10.3233/SW-160213. https://doi.org/10.3233/SW-160213.

Hickey, Thomas B., and Jenny A. Toves. 2014. "Managing Ambiguity In VIAF." *DLib Magazoine* 20 (7/8). https://doi.org/doi:10.1045/july2014-hickey.

Hyvönen, Eero. 2022. "Digital Humanities on the Semantic Web: Sampo Model and Portal Series." *Semantic Web* 14 (4): 729–744. https://doi.org/10.3233/SW-223034.

Hyvönen, Eero. 2024. "How to Create a National Cross-domain Ontology and Linked Data Infrastructure and Use It on the Semantic Web." DOI: 10.3233/SW-243468, *Semantic Web,* https://doi.org/10.3233/SW-243468.

Hyvönen, Eero, Erkki Heino, Petri Leskinen, Esko Ikkala, Mikko Koho, Minna Tamper, Jouni Tuominen, and Eetu Mäkelä. 2016. "WarSampo Data Service and Semantic Portal for Publishing Linked Open Data about the Second World War History." In *The Semantic Web – Latest Advances and New Domains (ESWC 2016),* edited by Harald Sack, Eva Blomqvist, Mathieu d'Aquin, Chiara Ghidini, Simone Paolo Ponzetto, and Christoph Lange, 758–773. Springer–Verlag, May. https://doi.org/10.1007/978-3-319-34129-3_46.

Hyvönen, Eero, and Heikki Rantala. 2021. "Knowledge-based Relational Search in Cultural Heritage Linked Data." *Digital Scholarship in the Humanities (DSH)* 16 (Supplement_2): ii155–ii164. https://doi.org/10.1093/llc/fqab042. https://doi.org/10.1093/llc/fqab042.

Hyvönen, Eero, Laura Sinikallio, Petri Leskinen, Senka Drobac, Rafael Leal, Matti La Mela, Jouni Tuominen, Henna Poikkimäki, and Heikki Rantala. 2024. "Publishing and Using Parliamentary Linked Data on the Semantic Web: ParliamentSampo System for Parliament of Finland." In print, *Semantic Web* (October). https://seco.cs.aalto.fi/publications/2024/hyvonen-et-al-ps-swj-2024.pdf.

Koho, Mikko, Esko Ikkala, Petri Leskinen, Minna Tamper, Jouni Tuominen, and Eero Hyvönen. 2021. "WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data." *Semantic Web* 12, no. 2 (January): 265–278. https://doi.org/10.3233/SW-200392. https://doi.org/10.3233/SW-200392.

Lehmann, Jens, Jörg Schüppel, and Sören Auer. 2007. "Discovering Unknown Connections—the DBpedia Relationship Finder." In *Proc. of the 1st Conference on Social Semantic Web (CSSW 2007),* 113:99–110. LNI. GI. http://subs.emis.de/LNI/Proceedings/Proceedings113/gi-proc-113-010.pdf.

Leskinen, Petri. 2024. "Modeling and Using Biographical Linked Data for Prosopographical Data Analysis." PhD diss., Aalto University, School of Science, Department of Computer Science, October. https://seco.cs.aalto.fi/publications/2024/leskinen-phd-2024.pdf.

Leskinen, Petri, Eero Hyvönen, and Jouni Tuominen. 2021. "Members of Parliament in Finland Knowledge Graph and Its Linked Open Data Service." In *Further with Knowledge Graphs. Proceedings of the 17th International Conference on Semantic Systems, 6-9 September 2021, Amsterdam, The Netherlands,* 255–269. IOS Press. https://doi.org/10.3233/SSW210049. https://doi.org/10.3233/SSW210049.

Rantala, Heikki, Eero Hyvönen, and Petri Leskinen. 2023. "Finding and explaining relations in a biographical knowledge graph based on life events: Case BiographySampo." In *Joint Proceedings of the ESWC 2023 Workshops and Tutorials co-located with 20th European Semantic Web Conference (ESWC 2023),* vol. 3443. CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3443/ESWC_2023_SEMMES_relations.pdf.

Rantala, Heikki, Petri Leskinen, Lilli Peura, and Eero Hyvönen. 2024. "Representing and searching associations in cultural heritage knowledge graphs using faceted search." In *SEMANTiCS 2024, 20th International Conference on Semantic Systems, proceedings.* In press. IOS Press, September. https://seco.cs.aalto.fi/publications/2024/rantala-et-al-searching-interesting-relations-2024.pdf.

Tartari, Gonzalo, and Aidan Hogan. 2018. "WiSP: Weighted Shortest Paths for RDF Graphs." In *Proceedings of VOILA 2018,* 37–52. CEUR Workshop Proceedings, vol. 2187.