BookSampo Fiction Literature Knowledge Graph Revisited: Building a Faceted Search Interface with Seamlessly Integrated Data-analytic Tools

Eero Hyvönen[1,2], Annastiina Ahola[1], and Esko Ikkala[1]

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

Abstract. *BookSampo – Finnish Fiction Literature on the Semantic Web* is a Linked Data (LD) service and portal deployed in 2011 by the Public Libraries of Finland. The portal has since then become a popular national web service with nearly 2 million annual users. However, BookSampo's user interface (UI) on top of the underlying knowledge graph (KG) in a SPARQL endpoint has been created using traditional search and data exploration methods, such as text search, and the full potential of the semantically rich KG has not been fully utilized. The data, covering detailed semantic metadata about all Finnish fiction literature and beyond, some 214 000 manifestations in total, is also interesting from a Digital Humanities (DH) research perspective. This paper presents first lessons learned and results in creating a new, additional UI for the BookSampo KG using faceted semantic search integrated seamlessly to data-analytic tools for DH research. At the same time, a Linked Open Data version (CC BY 4.0) of BookSampo is being created on materials whose copyright allows this.

Keywords: Semantic Web, Digital Libraries, Linked Data, User Interfaces, Portals

1 BookSampo Data and Portal: From Research to Practise

BookSampo[3] provides information on virtually all fiction literature published in Finland since mid 19th century. Its contents are based rich semantic descriptions of books and theirs contexts using Linked Data (LD) that originates from multiple heterogeneous data sources. BookSampo is an application instance of the more general "Sampo Model"[4] for LD publishing and series of semantic portals in use[5] in Finland and beyond [11].

BookSampo is used by library users and librarians for finding literary works of interest and related contextual information. The current original Drupal-based user interface (UI) in use since 2011 provides traditional text search engines for finding records and

[3] See project research homepage https://seco.cs.aalto.fi/applications/kirjasampo/
[4] The model is called "Sampo" according to the Finnish epic Kalevala, where Sampo is a mythical machine giving riches and fortune to its holder, a kind of ancient metaphor of technology according to the most common interpretation of the concept.
[5] See https://seco.cs.aalto.fi/applications/sampo/ for Sampos, links, videos, and publications.

then related contents as links for data exploration. However, the full potential of the data for searching, exploration, and for data analytic research has not been used as in other later Sampo systems. This paper presents a case study on developing a semantic UI for BookSampo, based on faceted search and seamlessly integrated data analytic tools, using the Sampo model and Sampo-UI framework [12]. We report lessons learned in applying the new declarative version of Sampo-UI available at Github[6] to a KG.

This paper first overviews the BookSampo KG as well as the Sampo principles and Sampo-UI framework to be used in the new UI design (Section 2). After this in Section 3, their application to the BookSampo KG is presented with examples of using the new demonstrator. In conclusion, contributions of the case study are summarized, related works discussed, and next steps ahead are outlined.

2 Sampo Model and Sampo-UI Framework

Data Model and Data BookSampo was developed as part of the larger system *Culture-Sampo – Finnish Culture on the Semantic Web* [9,17], a result of the national FinnONTO research initiative (2003–2012)[7] [8]. BookSampo was deployed in 2011 and has been maintained by the Finnish Public Libraries[8] as a popular national web service[9]. The original BookSampo dataset, described in detail in [20,16], contained nearly $400\,000$ instances of classes for literary works, authors, book covers, reviews, awards, images, and movies, over 3 million triples in total. The linked data was transformed from legacy library databases, and were then annotated and enriched manually by dozens of librarians in a Web 2.0 fashion. The data is constantly updated by metadata about new publications that are aligned with the annotation ontologies in a semi-automatic process. Today, the size of the triple store is already $8\,740\,000$.

Table 1. Sampo Model Principles P1–P6

P1	Support collaborative data creation and publishing
P2	Use a shared open ontology infrastructure
P3	Make clear distinction between the LOD service and the user interface (UI)
P4	Provide multiple perspectives to the same data
P5	Standardize portal usage by a simple filter-analyze two-step cycle
P6	Support data analysis and knowledge discovery in addition to data exploration

Sampo Model BookSampo is based on the Sampo Model [11], an informal collection of principles for LOD publishing and designing semantic portals listed in Table 1. Principles P1–P3 are used for developing data services: The model is based on the idea of collaborative content creation (P1) from local data silos, based on a shared ontology infrastructure (P2). The model argues for separating the underlying Linked Data

[6] See documentation at: https://github.com/SemanticComputing/sampo-ui
[7] FinnONTO homepage: https://seco.cs.aalto.fi/projects/finnonto/
[8] https://kirjastot.fi
[9] Available at: http://kirjasampo.fi

service *completely* from the UI via a SPARQL API. This arguable simplifies the portal architecture and the data service can be opened for data analysis research in Digital Humanities. The goal of P4–P6 is to "standardize" the UI logic so that the portals are easier to use [12]. Principle P4 articulates the idea of providing different thematic *application perspectives* by re-using the data service. They are provided on the landing page of the Sampo portal or as separate applications by third parties. The application perspectives are used by a two-step cycle for research (P5): First the focus of interest, the target group, is filtered out using faceted semantic search [7,23,24]. Second, the target group is visualized or analyzed by using ready-to-use data analytic tools of the application perspectives. The Sampo model aims not only at data publishing with search and data exploration [18] but also to data analysis and knowledge discovery with seamlessly integrated tooling for finding, analysing, and even solving research problems in interactive ways (P6) using AI [10].

Sampo-UI [12] is a new software framework for developing user interfaces for semantic portals, based on the principles P3 and P4–P6. The framework makes it possible to create highly customizable, user-friendly, and responsive user interfaces using current state-of-the-art JavaScript libraries and data from SPARQL endpoints, while saving substantial coding effort. Sampo-UI is available on GitHub under the open MIT License and has been utilized in alla Sampo portals since NameSampo [13] for toponomastic research that was published in 2019. The tool has also been used in external development projects, e.g., in Finland, Norway, and the U.S.

3 Sampo-UI Interface for BookSampo Knowledge Graph

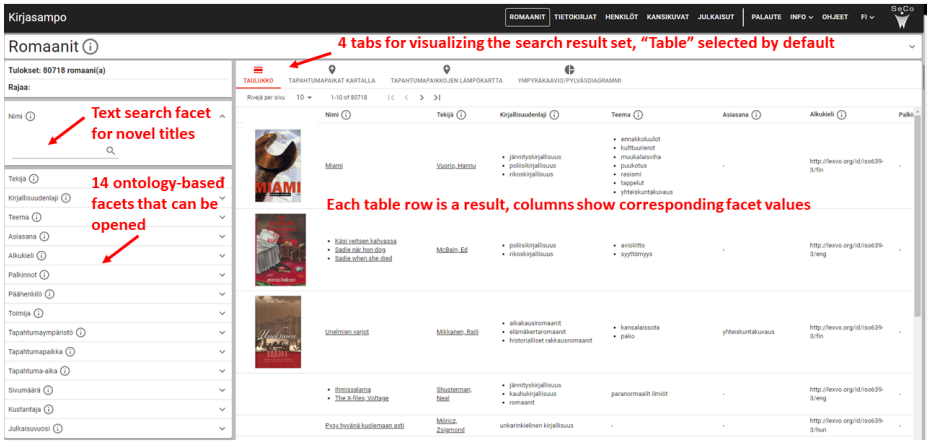

Fig. 1. Application perspective for searching, visualizing, and analyzing novels

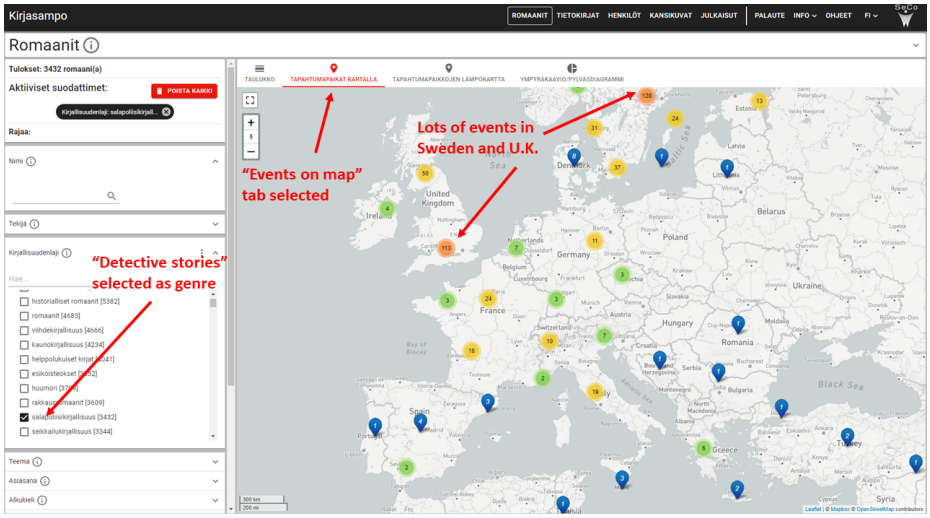

Fig. 2. Detective stories on a map. By clicking a place marker, related novels are found.

The aim of developing a new UI for BookSampo KG is to provide the user with facilities for 1) faceted semantic search and browsing as well as for 2) data-analytic tools and visualizations of the KG. The UI consists of a landing page with five thematic perspectives for searching and analyzing instances of 1) Novels (80 718 pcs), 2) Nonfiction books (1956), 3) Persons (62 207), 4) Book covers (112 971), and 5) Different manifestations of the books (213 877). Each application perspective includes three parts, as illustrated in Fig.1 for the Novels application perspective: 1) A faceted search engine on the left, based on the annotation ontologies of the KG; here 15 facets are used: Book title text search, Authors, Genres, Themes, Keywords, Original languages, Prizes, Main persons, Actors, Narrative environments, Places, Times, Number of pages, Publishers, Years of publishing. 2) A paginated results table view on the right, where the search results are shown by default as table rows with facet values as columns. 3) Tabs on the top of the search results for other data-analytic views of the results. Clicking on a result lets the user to see and navigate the "homepage" of the result instance that aggregates all data, images, and links, both internal and external, regarding the instance. Also instance pages may have a set of tabs to be used for data-analytic views of the instance and for exporting the data.

In the Novels view there are 4 visualizations: Table of results (default), Events on map, Heat map of places, and Pie charts for data-analysis. In Fig. 2 the user has filtered "detectives stories" using the Genre facet and selected the tab "Events on map".

The same UI logic is used in the other perspectives, too. For example, in Fig. 3 female authors were selected in the Persons perspective and their migrations from the place of birth (blue end of an arrow) to the place death (red end of an arrow) is visualized. By clicking on an arrow, the homepages of the corresponding persons are found, such as Doris Lessing based on the arrow from Iran to London.

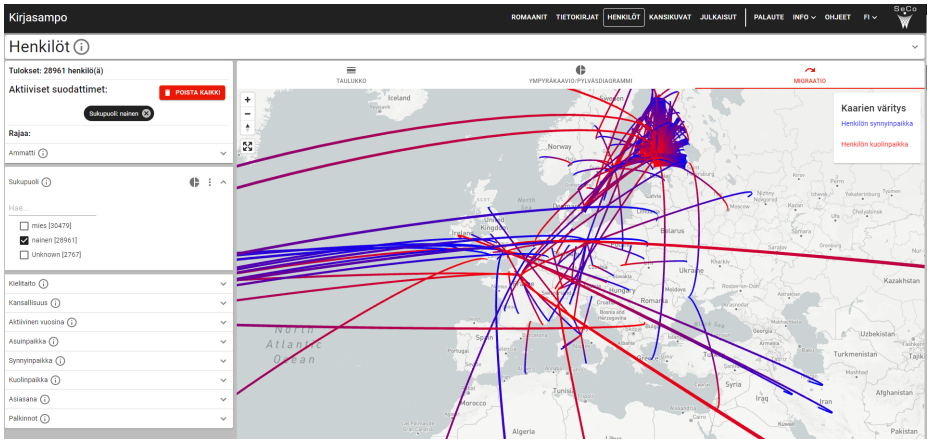

Fig. 3. Migrations of female authors visualized in Persons application perspective. Clickable arcs point from places of birth (blue end) to places of death (red end) of authors.

The visualizations are implemented by using components included in Sampo-UI or as custom made ones that have been built with visualization libraries, such as ApexCharts[10] and deck.gl[11].

4 Discussion, Contributions and Related Works

About Data and Its Quality Moving from text search to faceted ontology-based search exposed challenges in the underlying data models and in the data quality. For example, pseudonyms of authors were modelled as separate instances of the class Person, which means that novels of a certain person cannot be found easily when using the Person facet. A distinction between persons and their roles needs to be made. As for data quality, the new faceted UI automatically highlights problems in data, such as wrong data types, and missing property values, labels, and language tags. Also duplicate instances of data were found. The good news is that the UI can now be used as an easy-to-use visual aid for detecting problems in the data to be corrected later in the project. To truly utilize the full potential of the data, the data would need to be cleaned up and completed.

A deeper UI challenge in using the data for research is that using structured linked data requires new kind data literacy [14] and source criticism[12] from the end user. In big datasets it may not be possible to check and correct all data manually. Furthermore, the linked data approach is based on using explicit classifications and ontologies for which different opinions may arise. In many cases, the underlying real world may be too complex to be modelled fully. For example, the place ontology covers centuries

[10] https://apexcharts.com/
[11] https://deck.gl/
[12] https://ranke2.uni.lu/define-dsc/#%20,%20Universit%C3%A9%20du%20Luxembourg

of places that change in time. For example, Finland was part of Sweden until 1809, then part of Russia until becoming independent in 1917, and after that some parts of her were annexed to the Soviet Union that became later the modern Russia. What the underlying data actually means is not always clear and issues of Big Data quality, such as incompleteness, veracity, skewness, uncertainty, fuzziness, and errors of data arise.

Re-using Sampo-UI for New Sampos The latest version of the Sampo-UI at Github follows declarative programming paradigm where the idea is to replace procedural descriptions by configurations of the UI components, such as the search facets based on SPARQL queries. A goal of our case study was the test how easy the new Sampo-UI version is to use for an external developer not familiar with the tool. Our experiment suggests that configuring the existing components is easy for creating the application perspectives, once one has implemented his/her first perspective. The work can be started by copying an earlier project and then starting to tweak and adapt its components for the new application case and SPARQL endpoint. Also reusing existing components for implementing the data-analytic tools in application perspective tabs and for visualizations on instance homepages was deemed easy. Sophisticated interfaces could be made with the ready-to use Sampo-UI components, and a first version of the demonstrator could be shown to the librarians in two weeks. Creating new components for the Sampo-UI framework would need more sophisticated skills and knowledge about Sampo-UI.

Related work Linked Data and ontologies have been used as a basis for publishing collections in libraries [5], museums [1,21,22], and archives [25,4]. Using linked (open) data as a basis for library data is advocated by major library organizations, such as IFLA[13] and OCLC [14], and several national and other libraries provide their collections as data in this form [19]. Linked data has been used in building infrastructures, such as ARIADNEplus [15] for archaeology, Linked Art[16] in the U.S., and in local efforts in Italy [3], the U.K. [15], Spain [6], and Finland [8] to list a few examples. CH and DH have become a major application domain for Linked Data technologies [2,26]. However, there has been less research and development work on how to provide the data to end users through intelligent UIs. The novelty of the BookSampo KG and its new UI demonstrator lays in the semantic richness of the underlying KG and the integration of data-analytic tooling for DH research with faceted semantic search and browsing, as suggested by the Sampo model [10,12,11]. The focus in related research on portal UIs has been usually on (explorative) search and browsing [18].

Acknowledgements Thanks to Matti Sarmela, Kaisa Hypén, and Tuomas Aitonurmi for providing the new version of BookSampo KG and for research collaborations. Our work was funded by the Aalto University and is also related to the EU project InTaVia[17].

[13] https://www.ifla.org/references/best-practice-for-national-bibliographic-agencies-in-a-digital-age/service-delivery/linked-open-data/

[14] https://www.oclc.org/research/areas/data-science/linkeddata/linked-data-overview.html

[15] https://ariadne-infrastructure.eu/

[16] https://linked.art/

[17] https://intavia.eu

References

1. Aroyo, L., Stash, N., Wang, Y., Gorgels, P., Rutledge, L.: CHIP demonstrator: Semantics-driven recommendations and museum tour generation. In: The Semantic Web, pp. 879–886. Springer (2007). https://doi.org/10.1007/978-3-540-76298-0_64
2. Bikakis, A., Hyvönen, E., Jean, S., Markhoff, B., Mosca, A.: Special issue on semantic web for cultural heritage. Semantic Web – Interoperability, Usability, Applicability **12**(2) (2021). https://doi.org/10.3233/SW-210425
3. Carriero, V.A., Gangemi, A., Mancinelli, M.L., Marinucci, L., Nuzzolese, A.G., Presutti, V., Veninata, C.: ArCo: The italian cultural heritage knowledge graph. In: The Semantic Web – ISWC 2019. pp. 36–52. Springer (2019). https://doi.org/10.1007/978-3-030-30796-7_3
4. Hallo, M., Luján-Mora, S., Maté, A., Trujillo, J.: Current state of linked data in digital libraries. Journal of Information Science **42**(2), 117–127 (2016). https://doi.org/10.1177/0165551515594729
5. Haslhofer, B., Isaac, A., Simon, R.: Knowledge graphs in the libraries and digital humanities domain. arXiv preprint arXiv:1803.03198 (2018)
6. Hernández, F., Rodrigo, L., Contreras, J., Carbone, F.: Building a cultural heritage ontology for Cantabria. In: Annual Conference of CIDOC. pp. 1–14 (2008), https://cidoc.mini.icom.museum/wp-content/uploads/sites/6/2018/12/64_papers.pdf
7. Hyvönen, E., Saarela, S., Viljanen, K.: Application of ontology-based techniques to view-based semantic search and browsing. In: Proceedings of the First European Semantic Web Symposium. Springer (2004)
8. Hyvönen, E., Viljanen, K., Tuominen, J., Seppälä, K.: Building a National Semantic Web Ontology and Ontology Service Infrastructure – The FinnONTO Approach. In: Proceedings of the ESWC 2008, Tenerife, Spain. pp. 95–109. Springer (2008)
9. 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. Archives & Museum Informatics, Toronto (2009)
10. Hyvönen, E.: Using the Semantic Web in Digital Humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. Semantic Web – Interoperability, Usability, Applicability **11**(1), 187–193 (2020)
11. Hyvönen, E.: Digital humanities on the semantic web: Sampo model and portal series. Semantic Web – Interoperability, Usability, Applicability (2022), http://semantic-web-journal.org/content/digital-humanities-semantic-web-sampo-model-and-portal-series, accepted
12. 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). https://doi.org/10.3233/SW-210428
13. Ikkala, E., Tuominen, J., Raunamaa, J., Aalto, T., Ainiala, T., Uusitalo, H., Hyvönen, E.: Namesampo: A linked open data infrastructure and workbench for toponomastic research. In: Proceedings of the 2nd ACM SIGSPATIAL Workshop on Geospatial Humanities. pp. 2:1–2:9. GeoHumanities'18, ACM, New York, NY, USA (November 2018). https://doi.org/10.1145/3282933.3282936
14. Koltay, T.: Data literacy for researchers and data librarians. Journal of Librarianship and Information Science **49**(1), 3–14 (2015). https://doi.org/10.1177/0961000615616450
15. Lei, Y., Lopez, V., Motta, E., Uren, V.: An infrastructure for semantic web portals. Journal of Web Engineering **6**(4), 283–308 (2007), https://journals.riverpublishers.com/index.php/JWE/article/view/4105

16. Mäkelä, E., Hypén, K., Hyvönen, E.: BookSampo—lessons learned in creating a semantic portal for fiction literature. In: Proc. of ISWC-2011, Bonn, Germany. Springer (2011)
17. 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)
18. Marchionini, G.: Exploratory search: from finding to understanding. Communications of the ACM **49**(4), 41–46 (2006). https://doi.org/10.1145/1121949.1121979
19. Mitchell, E.T.: Library linked data: early activity and development. ALA TechSource Chicago, IL (2016)
20. Mäkelä, E., Hypén, K., Hyvönen, E.: Fiction literature as linked open data – the booksampo dataset (2013). https://doi.org/10.3233/SW-120093
21. Schreiber, G., Amin, A., Aroyo, L., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Omelayenko, B., van Osenbruggen, J., Tordai, A., Wielemaker, J., Wielinga, B.: Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator. Journal of Web Semantics **6**(4), 243–249 (2008). https://doi.org/10.1016/j.websem.2008.08.001
22. Szekely, P., Knoblock, C.A., Yang, F., Fink, E.E., Gupta, S., Allen, R., Goodlander, G.: Publishing the data of the Smithsonian American Art Museum to the linked data cloud. International Journal of Humanities and Arts Computing **8**(supplement), 152–166 (2014), http://usc-isi-i2.github.io/papers/szekely14-ijhac.pdf
23. Tunkelang, D.: Faceted search. Morgan & Claypool Publishers, CA, USA (2009)
24. Tzitzikas, Y., Manolis, N., Papadakos, P.: Faceted exploration of RDF/S datasets: a survey. Journal of Intelligent Information Systems **48**(2), 329–364 (2017). https://doi.org/10.1007/s10844-016-0413-8
25. Van Hooland, S., Verborgh, R.: Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata. Facet Publishing (2014). https://doi.org/10.1080/00048623.2016.1162277
26. Zeng, M., Sula, C., Gracy, K., Hyvönen, E., Lima, V.M.A.: JASIST special issue on digital humanities (DH). Journal of the Association for Information Science and Technology (JASIST) pp. 1–5 (2021). https://doi.org/10.1002/asi.24584