"Sampo" Model and Semantic Portals for Digital Humanities on the Semantic Web

Eero Hyvönen

University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG), and
Aalto University, Semantic Computing Research Group (SeCo)
eero.hyvonen@aalto.fi

Abstract. This paper presents the vision and longstanding work in Finland on creating a national Cultural Heritage ontology infrastructure and semantic portals based on Linked Data on the Semantic Web. In particular, the "Sampo" series of semantic portals is considered, including CultureSampo (2009), TravelSampo (2011), BookSampo (2011), WarSampo (2015), BiographySampo (2018), NameSampo (2019), WarWictimSampo (2019), FindSampo (2019), MMM (2020), LawSampo (2020), AcademySampo (2020), and ParliamentSampo (2022). They all share the "Sampo model" for publishing Cultural Heritage content the Semantic Web that typically involves three components: 1) A "business model" for harmonizing, aggregating, and publishing heterogeneous, distributed contents based on a shared ontology infrastructure. 2) An approach to interface design, where the data can be re-used and accessed independently from multiple application perspectives, while the data resides in a single SPARQL endpoint. 3) A two-step model for accessing and analyzing the data where the focus of interest is first filtered out using faceted semantic search, and then visualized or analyzed by ready-to-use Digital Humanities tools of the portal. This model has been proven useful in practise: Sampo portals have attracted lots users from tens of thousands to millions depending on the Sampo. It is argued that the next step ahead could be portals for serendipitous knowledge discovery where the tools, based on AI techniques, are able to find *automatically* serendipitous, "interesting" phenomena and research questions in the data, and even solve problems with explanations.

Keywords: Semantic Web, Semantic Portal, Digital Humanities.

1 Vision: Cultural Heritage on the Semantic Web

A fundamental semantic problem in publishing and using Cultural Heritage (CH) data on the Web is how to make the heterogeneous CH contents semantically interoperable, so that they can be searched, interlinked, and presented in a harmonized way across the boundaries of the datasets and data silos. The problem is related to the way CH content is created: the data is collected, maintained, and published by different museums, libraries, archives, and other actors using their own standards and best practices that may not be compatible with each other. Semantic Web (SW) technologies [2] and Linked Data [1,4] are a promising approach for addressing the problems of semantic interoperability in a distributed content creation environment. Based on computer

"understandable" linked big data, intelligent applications for Digital Humanities (DH) researchers and the public can be created.

This paper presents an approach for making these promises of Linked (Open) Data and the Semantic Web to come true: the "Sampo model" and its application in practise in creating a series of semantic portals in use in Finland. The model is based on what is today called FAIR principles[1]. In the following, the Sampo model is first outlined and after that the Sampo-series of semantic portals is presented. In conclusion, a vision ahead towards a next "third generation" of semantic portals based on serendipitous knowledge discovery and Artificial Intelligence is presented, based on [6].

2 Sampo Model for Publishing and Using Linked Data

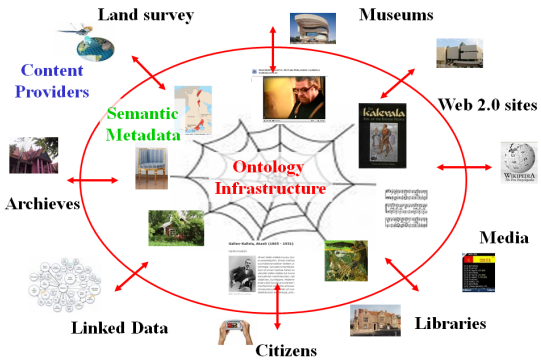

Fig. 1. Publishing heterogeneous distributed data in the Semantic Web

The "Sampo model" includes three components related to 1) the model for creating and publishing heterogeneous, distributed Linked Data, 2) to providing the end user with multiple application perspectives to the contents, and 3) how the application perspectives can be used in two basic steps. I call this model "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[2]. The idea of collaborative content creation by data linking is a fundamental idea behind the Linked Data movement[3] and has been developed also in various other settings, e.g., in ResearchSpace[4].

1. Data creation and publishing model. The ideas of the Semantic Web and Linked Data can be applied to address the problems of semantic data interoperability and dis-

[1] Findable, Accessible, Interoperable, and Re-usable, cf., https://www.go-fair.org/fair-principles/.
[2] https://en.wikipedia.org/wiki/Sampo
[3] http://linkeddata.org/
[4] https://www.researchspace.org/

tributed content creation at the same time, as depicted in Fig. 1. Here the publication system is illustrated by a circle. A shared semantic ontology infrastructure is situated in the middle. It includes mutually aligned shared domain ontologies and core metadata, modeled by using SW standards[5]. If content providers outside of the circle provide the system with metadata about their contents, the data is automatically linked and enriched with each other and forms a knowledge graph represented using RDF[6]. For example, if metadata about a painting created by Picasso comes from an art museum, it can be enriched (linked) with, e.g., biographies from Wikipedia and other sources, photos taken of Picasso, information about his wives, books in a library describing his works of art, related exhibitions open in museums, and so on. At the same time, the contents of any organization in the portal having Picasso related material get enriched by the metadata of the new artwork entered in the system. This is a win-win business model for everybody to join; collaboration pays off. The knowledge graph can be published as a data service in a SPARQL end-point using the principles of Linked Data [1].

2. Multiple perspective interface design. On top of the data service different applications can be created by re-using the data service, without modifying the data. For example, in WarSampo [7] the data about the Second World War can be accessed from nine points of view: historical events, people, units, places, articles, death records, photographs, cemeteries, and prisoners of war. In Sampo portals the application perspectives are provided on the landing page of the system. By selecting one of them the corresponding application is opened.

2. Filter-analyze two-step usage cycle. In many Sampos, the application perspectives can be used by a two-step cycle for research: First the focus of interest, the target group, is filtered out using faceted semantic search [18]. Second, the target group is visualized or analyzed by using ready-to-use DH tools of the application perspectives. For example, in BiographySampo [9] a group of people, such as the clergy of the 19th century Graph Duchy of Finland, can be filtered out first. After this, the life charts of the priests from places of birth to death can be visualized for analyzing their mobility, their mutual network be visualized, various statistics of the group be viewed, and so on.

3 Sampo Series of Semantic Portals

To develop, test, and demonstrate the model, a series of "Sampo" portals have been created and are in use on the Semantic Web in Finland. These living lab prototypes and applications have been created as part of research projects at the Semantic Computing Research Group (SeCo) active at Aalto University and the University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG), and are based on collaborations with a large network of Finnish memory and other organizations as data providers and cultural heritage domain experts. The systems are examples of utilizing a national level FinnONTO ontology and Linked Open Data infrastructure [3] that has been developed in conjunction with the portals:

[5] https://www.w3.org/standards/semanticweb/
[6] https://www.w3.org/RDF/

1. **CultureSampo – Finnish Culture on the Semantic Web 2.0**[7] (online since 2009) [5,14], demonstrates how CH content of tens of different kinds can enrich each other, including a semantic model of the Kalevala epic narrative at the center.
2. **BookSampo – Finnish Fiction Literature on the Semantic Web**[8] (online since 2011) [13] publishes metadata about virtually all Finnish fiction literature as a knowledge graph on top of which a portal was created. BookSampo data was originally part of CultureSampo. BookSampo is today maintained by the Public Libraries of Finland and is used by ca. 2 million users in a year.
3. **TravelSampo – Mobile Contextualized Services of Cultural Tourism**[9] (published in 2011) [15] pioneered the idea of providing cultural content to mobile travelers in a personalized and real world context.
4. **WarSampo – Finnish World War II on the Semantic Web**[10] (online since 2015) [7] is a popular Finnish service that has had 570 000 users . It provides information about the ca. 100 000 casualties and significant soldiers of the WW2 in Finland and various datasets, such as 160 000 photographs from the fronts, war diaries, maps etc. A key idea in WarSampo is to reassemble the life stories of the soldiers based on data linking from different data sources. See the online video "WarSampo"[11] illustrating the system.
5. **BiographySampo – Finnish Biographies on the Semantic Web**[12] (online since 2018) [9] is yet another popular service with tens of thousands of users. It is based on mining out a large knowledge graph (over 120 million triples) from ca. 13 100 Finnish biographies of the Finnish Literature Society, authored by 1000 scholars. The data is interlinked and enriched internally and by some 16 external datasources. See the online video "BiographySampo – Artificial Intelligence Reading Biographies for the Semantic Web"[13] for the underlying vision and the actual system.
6. **NameSampo – A Linked Open Data Infrastructure and Workbench for Toponomastic Research**[14] (online since 2019) [11] publishes data about over 2 million place names and places in Finland with old maps. It soon attracted tens of thousands of users on the Web. The data originates from the Institute of Languages of Finland, National Survey of Finland, Getty Thesaurus of Geographical Names, and various map services, including historical maps.

In addition, there are several new Sampos to be published in the near future: **FindSampo**[15] (on archaeology and citizen science) [19], **WarVictimSampo**[16] (on Finnish wars 1914–1922) [17], **AcademySampo**[17] (on Finnish academic people 1640–1899)

[7] https://seco.cs.aalto.fi/applications/kulttuurisampo/
[8] https://seco.cs.aalto.fi/applications/kirjasampo/
[9] https://seco.cs.aalto.fi/applications/travelsampo/
[10] https://seco.cs.aalto.fi/projects/sotasampo/en/
[11] https://vimeo.com/212249404
[12] https://seco.cs.aalto.fi/projects/biografiasampo/en/
[13] https://vimeo.com/328419960
[14] https://seco.cs.aalto.fi/projects/nimisampo/en/
[15] https://seco.cs.aalto.fi/projects/sualt/
[16] https://seco.cs.aalto.fi/projects/sotasurmat-1914-1922/
[17] https://seco.cs.aalto.fi/projects/yo-matrikkelit/

[12], and **LawSampo**[18] (on Finnish legislation and case law) [10]. Also the **MMM** portal [8], a result of the international Mapping Manuscript Migrations[19] project is based on the Sampo model, and work on developing **ParliamentSampo**[20] (on open data of the Parliament of Finland) has started.

4 Towards Knowledge Discovery and Artificial Intelligence

Early Sampos and current state-of-the-art CH portals, such as Europeana[21] and Digital Public Library of America[22], have focused on data aggregation, enrichment, search, and exploration of data. However, there are also systems for not only searching and browsing but also inspecting the data using visualizations and data-analysis. Visualizations were used already in CultureSampo, and in WarSampo data-analysis of casualties of war is possible. In BiographySampo and NameSampo the idea on providing data analytic tooling for DH researchers is already the main focus, and semantic search is seen more like a filtering phase of the data so that data analytic tools can be focused and applied on selected target data. Searching and browsing are only tools among others.

What is still largely missing in the DH methodology and tools in semantic portals is the next conceptual level of automatic knowledge discovery and Artificial Intelligence [16]. Why not create DH tools that are able not only to present the data to the human researcher in useful ways but also to 1) find DH research problems, 2) solve them *automatically by themselves*, and 3) also explain the reasoning or solution to the researcher? Artificial Intelligence techniques would also be useful when creating and enriching the knowledge graph underlying a semantic portal. First steps towards these goals have been taken in the BiographySampo where the underlying knowledge graph has been used for finding and explaining serendipitous semantic connections between places and persons to the end-user [9]. This vision is developed in more detail in [6].

Acknowledgments Tens of people have been working in developing the Sampo series, funded by ca. 50 organizations in 2003–2020 in Finland. The sites referred to in the footnotes contain full sets of publications online related to the systems, authored by the project members, as well as links to data, data services, and software.

References

1. 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/
2. Hitzler, P., Krötzsch, M., Rudolph, S.: Foundations of Semantic Web technologies. Springer–Verlag (2010).
3. 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–Verlag (2008).

[18] https://seco.cs.aalto.fi/projects/lawlod/
[19] https://seco.cs.aalto.fi/projects/mmm/
[20] https://seco.cs.aalto.fi/projects/semparl/en/
[21] http://europeana.eu
[22] https://dp.la/

4. Hyvönen, E.: Publishing and using cultural heritage linked data on the Semantic Web. Morgan & Claypool, Palo Alto, California (October 2012).
5. 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).
6. 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).
7. Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J., Mäkelä, E.: 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). pp. 758–773. Springer–Verlag (2016).
8. Hyvönen, E., Ikkala, E., Tuominen, J., Koho, M., Burrows, T., Ransom, L., Wijsman, H.: A linked open data service and portal for pre-modern manuscript research. CEUR Workshop Proceedings, Vol-2364 (2019).
9. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., Keravuori, K.: BiographySampo – Publishing and enriching biographies on the Semantic Web for digital humanities research. In: Proceedings of the 16th Extended Semantic Web Conference (ESWC 2019). pp. 574–589. Springer–Verlag (2019).
10. Hyvönen, E., Tamper, M., Ikkala, E., Sarsa, S., Oksanen, A., Tuominen, J., Hietanen, A.: LawSampo: A semantic portal on a linked open data service for Finnish legislation and case law (2019), white paper, https://seco.cs.aalto.fi/publications/2019/hyvonen-et-al-ls.pdf
11. 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, http://doi.acm.org/10.1145/3282933.3282936
12. Leskinen, P., Hyvönen, E.: Linked open data service about historical Finnish academic people in 1640–1899. In: Proc. of the Digital Humanities in the Nordic Countries (DHN 2020). CEUR WS Proceedings (2020), forth-coming
13. 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–Verlag (2011).
14. Mäkelä, E., Ruotsalo, T., Hyvönen: How to deal with massively heterogeneous cultural heritage data—lessons learned in CultureSampo. Semantic Web **3**(1), 85–109 (2012).
15. Mäkelä, E., Lindblad, A., Väätäinen, J., Alatalo, R., Suominen, O., Hyvönen, E.: Discovering places of interest through direct and indirect associations in heterogeneous sources – the TravelSampo system. In: Terra Cognita 2011: Foundations, Technologies and Applications of the Geospatial Web. CEUR Workshop Proceedings, Vol-798 (2011).
16. Pazzani, M.J.: Knowledge discovery from data? IEEE Intelligent Systems **15**, 10–13 (2000)
17. Rantala, H., Jokipii, I., Koho, M., Ikkala, E., Tuominen, J., Hyvönen, E.: Building a linked open data portal of war victims in Finland 1914–1922. In: Proc. of the Digital Humanities in the Nordic Countries (DHN 2020). CEUR WS Proceedings (2020), forth-coming.
18. Tunkelang, D.: Faceted search. Morgan & Claypool Publishers, CA, USA (2009).
19. Wessman, A., Thomas, S., Rohiola, V., Koho, M., Ikkala, E., Tuominen, J., Hyvönen, E., Kuitunen, J., Parviainen, H., Niukkanen, M.: Citizen science in archaeology: Developing a collaborative web service for archaeological finds in Finland. In: Jameson, J., Musteata, S. (eds.) Transforming Heritage Practice in the 21st Century: Contributions from Community Archaeology, pp. 337–352. Springer–Verlag (2019).