Linked Data in Use: Sampo Portals on the Semantic Web

Eero Hyvönen

[The Forcing of Sampo, a painting by A. Gallen-Kallela (1893). National Gallery of Finland.]

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 technologies and Linked Data are a promising approach for addressing the problems of semantic interoperability in a distributed content creation environment. The Semantic Web (SW) can be seen as a new layer of (meta)data being build inside the Web[1]. The methodology for representing metadata and ontological concepts on the Web is based on a simple data model for representing knowledge graphs. On a global WWW scale, the Semantic Web forms a Giant Global Graph (GGG, Web of Data) of connected data resources. The GGG can be used and browsed in ways analogous to the WWW, but while the WWW links associated web pages with each other for human use, the GGG links the underlying concepts and data resources together. For example, the GGG may tell that ducks are birds, and that Donald is an instance of a duck, and therefore also a bird, while the related WWW pages may constitute linked web documents about Donald Duck.
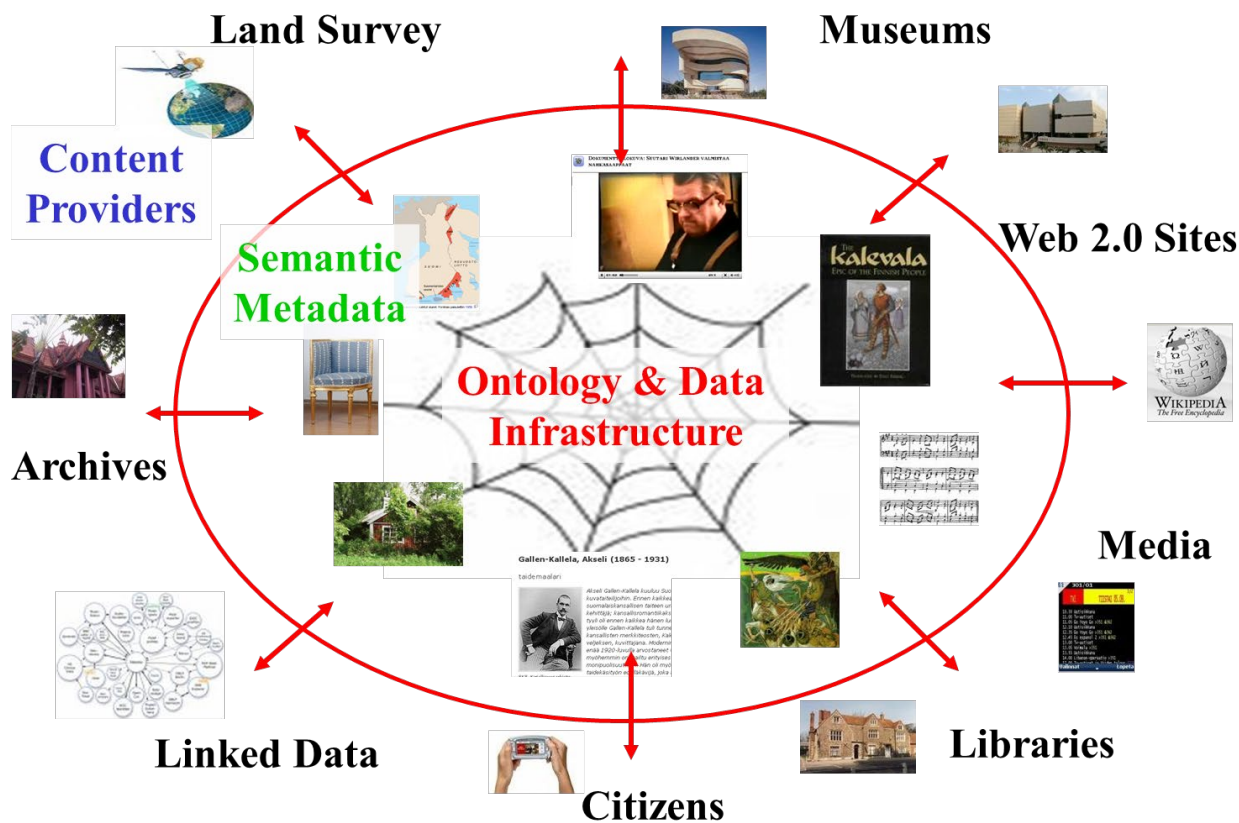
Sampo Model for Publishing Linked Data

Figure 1. Sampo model for publishing Linked Data in an interoperable way in a distributed environment[2].

The ideas of the Semantic Web and Linked Data can be applied to address the problems of semantic data interoperability and distributed content creation at the same time, as depicted in Figure 1. Here the publication system is illustrated by a circle. A shared semantic ontology infrastructure is situated in the middle. It includes mutually aligned metadata and shared domain ontologies, modeled by using SW standards. If content providers outside of the circle provide the system with metadata about CH, the data is automatically linked and enriched with each other and forms a GGG. 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. 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[3].

Linked Data at Work: "Sampo" Series of Semantic Portals

To test and demonstrate the idea, 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 Linked Open Data infrastructure[4] that has been developed in conjunction with the portals.

1. CultureSampo – Finnish Culture on the Semantic Web 2.0[5] (published in 2009), demonstrated 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[6] (online since 2011) published metadata about virtually all Finnish fiction books as a knowledge graph on top of which a portal was created, used today by ca. 2 million users in a year. BookSampo data was part of CultureSampo.
3. TravelSampo – Mobile Contextualized Services of Cultural Tourism[7] (published in 2011) pioneered the idea of providing cultural content to mobile travelers in real world context.
4. WarSampo – Finnish World War II on the Sematic Web[8] (published in 2015 and 2017) is a popular Finnish service of ca. 230 000 annual users providing information about the ca. 100 000 casualties and significant soldiers of the WW2 in Finland. A key idea in WarSampo is to reassemble their life stories based on data linking from different data sources. See the online video "WarSampo"[9] illustrating the system.
5. BiographySampo – Finnish Biographies on the Semantic Web[10] (published in 2018) is yet another popular service with thousands of users. It is based on mining out a large knowledge graph (over 100 million connections) from ca. 13 000 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"[11] for the underlying vision and the actual system.
6. NameSampo – A Linked Open Data Infrastructure and Workbench for Toponomastic Research[12] (published in 2019) 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.

In addition, two new sampos are under development: LawSampo for publishing Finnish legislation and law, based on data from the Ministry of Justice, and FindSampo for archaeological finds by the citizens, based on data from the National Heritage Agency of Finland and other sources.

Eero Hyvönen[13] is director of the Helsinki Centre for Digital Humanities (HELDIG) at the University of Helsinki and professor of semantic media technology at the Aalto University. He has been directing the research on Semantic Web infrastructures and applications of the Semantic Computing Research Group (SeCo) at the University of Helsinki, HELDIG, and Aalto University since 2002. Eero Hyvönen has published over 400 research articles and books and received several national and international awards. He acts in the editorial boards of Semantic Web - Interoperability, Usability, Applicability, Semantic Computing, International Journal of Metadata, Semantics, and Ontologies, and International Journal on Semantic Web and Information Systems, and has co-chaired and acted in the program committees of tens of major international conferences.

---

[1] This work is coordinated internationally by the World Wide Web consortium W3C: https://www.w3.org/standards/semanticweb/
[2] Eero Hyvönen: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Morgan & Claypool, Palo Alto, California, 2012.
[3] https://en.wikipedia.org/wiki/Sampo
[4] This infrastructure work started in 2003 as the national Finnish ontology project FinnONTO (https://seco.cs.aalto.fi/projects/finnonto/) and continues in HELDIG and Aalto University as the initiative Linked Open Data Infrastructure for Digital Humanities: https://seco.cs.aalto.fi/projects/lodi4dh/

[5] https://seco.cs.aalto.fi/applications/kulttuurisampo/
[6] https://seco.cs.aalto.fi/applications/kirjasampo/
[7] https://seco.cs.aalto.fi/applications/travelsampo/
[8] https://seco.cs.aalto.fi/projects/sotasampo/en/
[9] https://vimeo.com/212249404
[10] https://seco.cs.aalto.fi/projects/biografiasampo/en/
[11] https://vimeo.com/328419960
[12] https://seco.cs.aalto.fi/projects/nimisampo/en/
[13] https://seco.cs.aalto.fi/u/eahyvone/