Invited talk to be published in the proceedings of:
VIII Encounter of Documentation Centres of Contemporary Art:
Open Linked Data and Integral Management of Information in Cultural Centres
Artium, Vitoria-Gasteiz, Spain, October 19-20, 2016.

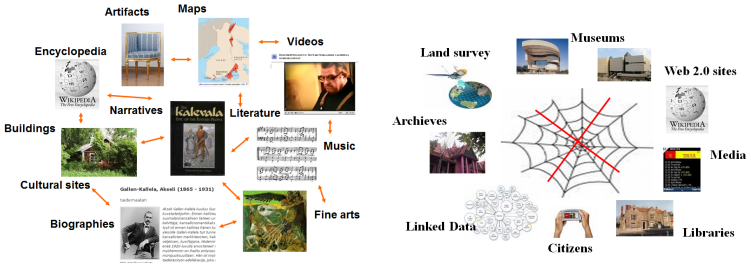
Cultural Heritage Linked Data on the Semantic Web: Three Case Studies Using the Sampo Model

Eero Hyvönen

Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, and
Semantic Computing Research Group (SeCo), Aalto University
http://heldig.fi, http://seco.cs.aalto.fi

Abstract. A major challenge in publishing linked Cultural Heritage (CH) collections on the web is interoperability. This is due to the heterogeneity of CH contents and the distributed content creation model where publishers focus on their own data with little consideration on the others' data. As a solution approach, the "Sampo" model is presented based on using domain independent modeling standards, on a model for aligning metadata models, and on sharing domain ontologies for populating the matadata models. The harmonized data is published for machines as a linked data service, to be used by applications for human users. To illustrate and evaluate the model, three online systems on the Web, Culture-Sampo, BookSampo, and WarSampo are presented.

1 Semantic Interoperability and Distributed Content Creation



(a) CH data is semantically heterogeneous and linked.

(b) CH content is produced by independent actors without coordination.

Fig. 1. Challenges for publishing CH collections on the Web.

Cultural Heritage (CH) data has many characteristic features: 1) The contents come in various forms, such as text documents, photos, audio tracks, videos, and collection item records. 2) The contents concern cross-domain topics, such as art, history, artifacts, and traditions. 3) The content is available in different languages. 4) The content is related and interpreted in terms of different cultures, such as religions or national traditions in the west and east. 5) The contents are often targeted to both laymen and

experts, young and old. At the same time, the data is semantically highly interlinked, as illustrated in Fig. 1(a).

As a result, a fundamental *semantic problem* in dealing with CH data is how to make the heterogeneous contents semantically *interoperable*, so that they can be searched, linked, and presented in a harmonized way across the boundaries of the datasets and data silos. In addition, there is also an *organizational problem* related to the way CH content is created: CH content 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, as illustrated in Fig. 1(b).

Semantic Web technologies[1] [2] are a promising approach for addressing the problems of semantic interoperability in a distributed content creation environment [4,7]. The Semantic Web (SW) can be seen as a new layer of *(meta)data* being build inside the Web. The methodology for representing metadata and ontological concepts on the Web is based on a simple data model: a directed labeled graph, i.e., a *semantic net*, based on the Resource Description Framework (RDF) model[2] and related standards. On a global WWW scale, the Semantic Web forms a *Giant Global Graph* (GGG) 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 a bird) while the related WWW pages may constitute a comics book about Donald Duck.

2 Sampo Model for Publishing Linked Data

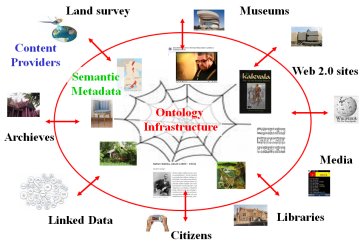


Fig. 2. Sampo model for Linked Data publishing is based on a shared ontology infrastructure in the middle.

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

[1] http://www.w3.org/standards/semanticweb/
[2] http://www.w3.ord/RDF/

time, as depicted in Fig. 2. 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 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 such a system; collaboration pays off. However to create a collaborative semantic portal for CH of this kind has a price, too. The linking can be established correctly only if unambiguous URI identifies are constantly used. Furthermore, following major *semantic agreements* are needed for interoperability:

1. **Domain neutral semantic model**. Agreement of using domain neutral Semantic Web standards of W3C, such as RDF(S), SKOS, and OWL.
2. **Metadata alignment model**. Agreement on using a metadata alignment model for harmonizing the different metadata models used by different partners.
3. **Shared domain ontologies**. Agreement of sharing domain ontologies (places, persons, etc.) whose concepts are used for populating the metadata models.

The semantic web standards (1) are used to harmonize everything cross-domain in the model, including the metadata models and domain ontologies. As for metadata models (2), especially two approaches are in use. First, Dublin Core and its dumb down principle[3] can be used for mapping different models (their elements) onto each other. Second, metadata of different forms can be mapped on a generic underlying ontological model of the world, such as CIDOC CRM[4]. The domain ontologies (3) include a set of cross-domain ontologies for general concepts (e.g., artifact types and materials), authorities (persons, groups, and organizations, places (current and historical), time periods, and events.

For publishing the content, separate *publishing services* are needed for the human end-users and the machines:

1. **Human users**. Semantic portal applications for end-users to search and browse data.
2. **Machine users**. Linked data services on which the semantic applications can be built upon. The enriched can data be re-used by everybody.

The three semantic agreements above constitute a kind of model for application domain infrastructures on the Semantic Web. Combining the infrastructure with the idea of decoupling the data services for machines form the applications for the human user creates a model for building collaborative Semantic Web applications. I call this whole

[3] http://dubincore.org
[4] http://cidoc-crm.org

the Sampo model. Th model has been developed and tested in a series of several practical case studies. In the following we overview three demonstrators based on the Sampo model: CulturesSampo (online since 2009), for cross-domain CH collections, BookSampo (online since 2011) for fiction literature, and WarSampo (online since 2015) for Second World War history. The systems are explained in terms the three semantic agreements and the two publishing services.

3 CultureSampo: Finnish Culture on the Semantic Web

CultureSampo[5] [8,10] is a publication system and a portal by which memory organizations (and even citizens) can publish their collections on the Semantic Web in a collaborative way. CultureSampo extends the earlier application "MuseumFinland—Finnish Museums on the Semantic Web"[6] [5], a system for publishing artifact collections on the Semantic Web, by supporting publication of different kind of cross-domain CH contents, both material and immaterial.

The Sampo model originates from CultureSampo that addressed two major challenges: 1) *Semantic challenges*. Cultural heritage content is semantically heterogeneous and interlinked. For example, the content may contain a person's narrative biography, works of art she created, places of interest where she lived in, Wikipedia articles or novels about or by the person, social connections to other persons, and events in the history that the person was related to. 2) *Organizational challenges*. Organizations and individual citizens create cultural heritage content independently from each other. This has led to a situation, where metadata produced by different organizations is usually incompatible with each other in terms of metadata schemas, vocabularies, and cataloging conventions.

CultureSampo was developed as part the FinnONTO project (2003-2010)[7], whose goal was build a national level semantic web content infrastructure in Finland [6] and demonstrate its usefulness in practical applications. CultureSampo fulfilled the three semantic agreements of the Sampo model in the following way: 1) The W3C Semantic Web recommendations were used as the domain independent framework. 2) Metadata model alignment—over 20 different models were used– was based on Dublin Core and its dumb down principle. 3) Shared domain ontologies: The basis of CultureSampo is the national FinnONTO infrastructure. It includes a collaboratively created system of cross-domain ontologies and related ontology services for utilizing them cost-efficiently as services. The ontologies and the services were published as the National Ontology Service ONKI[8] that is now maintained by the National Library under the name Finto.fi[9].

As for the publishing model, there is a semantic portal[10] for human users on the Web, based on a separate triple store service. When CultureSampo was created, suit-

[5] http://seco.cs.aalto.fi/applications/kulttuurisampo/
[6] http://www.museosuomi.fi
[7] http://www.seco.tkk.fi/projects/finnonto/
[8] http://www.onki.fi
[9] http://finto.fi/
[10] http://www.kulttuurisampo.fi

able SPARQL endpoint systems were not available, and an in-house triple store service based on Jena[11] was developed for the underlying data service.

4 BookSampo: Fiction literature on the Semantic Web

The role of public libraries as a provider of factual knowledge has diminished, but their role as a source for fiction literature remains strong [12]. The nature of fiction necessitates a departure from old library indexing traditions. Classifying books e.g. by their genre and by shelf location alone is not enough: studies on user needs in fiction literature show that satisfactory fiction literature indexing systems have to be able to also model and store the rich connections between fiction literature works, their authors, and their surrounding cultural context [11].

BookSampo system provides public library customers with a semantic information portal for finding and browsing rich interlinked information about fiction, thus completing services of traditional library catalog systems [9]. BookSampo is based on the Sampo model in the following way: 1) Domain neutral semantic model: the W3C Semantic Web recommendations were used. 2) Metadata model alignment: BookSampo is based on a version of the emerging FRBR-conceptual metadata model for bibliographi records[12] with Dublin Core. 3) Shared domain ontologies: As in CultureSampo, the basis of BookSampo is the national FinnONTO infrastructure.

The BookSampo dataset provides information as linked data on virtually all fiction literature published in Finland since mid-19th century. The dataset contains rich semantic descriptions, originally nearly 400,000 instances of data (subject URIs), including literary works, authors, book covers, reviews, awards, images, and movies, over 3 million triples in total. The metadata was transformed into RDF from legacy library databases, then annotated and enriched manually by dozens of librarians in a Web 2.0 fashion in Finnish public libraries, and is constantly updated as new books and related data are published. The data can be used for Digital Humanities research, too, e.g. to answer complex questions, such as what topics should one write about, if one wants to get a literary award (based on statistics).

BookSampo was developed in a joint venture of the Libraries.fi association of the Finnish public libraries[13] and Semantic Web researchers[14] at Aalto University and the University of Helsinki, as part of the national FinnONTO programme. The BookSampo semantic portal[15] was published 2011. It was provided first by the FinnONTO research project, but is today maintained independently by the Finnish public libraries. The portal had over 1.6 million visits in 2016 and is based on an underlying SPARQL data service of Linked Open Data.

[11] https://jena.apache.org
[12] http://www.ifla.org/frbr
[13] http://libraries.fi
[14] http://seco.cs.aalto.fi/applications/kirjasampo/
[15] http://www.kirjasampo.fi

5 WarSampo: Second World War on the Semantic Web

Many websites publish information about the Second World War (WW2), the largest global tragedy in human history[16]. Such information is of great interest not only to historians but to potentially hundreds of millions of citizens globally whose relatives participated in the war actions, creating a shared trauma all over the world. However, WW2 information on the web is typically meant for human consumption only, and there are hardly any web sites that serve *machine-readable data* about the WW2 for digital humanists [3,1] and end-user applications to use. It is our belief that by making war data more accessible, our understanding of the reality of the war improves, which not only advances understanding of the past but also promotes peace in the future.

"WarSampo—Second World War of the Semantic Web"[17] addresses these challenges. The idea is to 1) initiate and foster large scale LOD publication of WW2 data from distributed, heterogeneous data silos and 2) demonstrate and suggest its use in applications and research. WarSampo system publishes collections of heterogeneous, distributed data about the Second World War on the Semantic Web. The systems is based on harmonizing massive datasets using event-based modeling, which makes it possible to enrich datasets semantically with each other's contents. The Sampo model is used as follows: 1) Domain neutral semantic model: the W3C Semantic Web recommendations and the ISO standard for CIDOC CRM were used. 2) Metadata model alignment: WarSampo is based on the CIDOC CRM model, where all data in different forms is transformed into a harmonizing event-based metadata ontology model. 3) Shared domain ontologies: a set of domain specific ontologies for historical persons, places, army units etc. was created in order to harmonize and interlink data.

WarSampo has two components: First, there is a Linked Open Data (LOD) service WarSampo Data[18] for Digital Humanities (DH) research and for creating applications related to war history. The data comes from 8 different major datasets from different organizations, totaling originally 7.6 million triples in a SPARQL endpoint. Second, a semantic WarSampo Portal[19] has been created to test and demonstrate the usability of the data service. The portal allows both historians and laymen to study war history and destinies of their family members in the war from 7 different interlinked application perspectives. Published in November 2015, the WarSampo Portal had well over 10,000 distinct visitors during the first three days, showing that the public has great interest in these kind of applications.

World war history makes a promising use case for Linked Data (LD) and the Sampo model because war data is by nature heterogeneous, distributed in different countries and organizations, and written in different languages. When an organization contributes to the WW2 LOD cloud with a piece of information, say a photograph, its description is automatically connected to related data, such as persons or places depicted. At the same time, the related pieces of information, provided by others, are enriched with links to

[16] http://ww2db.com, http://www.world-war-2.info, different Wikipedias, etc.
[17] http://seco.cs.aalto.fi/projects/sotasampo/en/
[18] http://www.ldf.fi/dataset/warsa/
[19] http://www.sotasampo.fi

the new data. WarSampo is to our best knowledge the first large scale system for serving and publishing WW2 LOD on the Semantic Web.

6 Discussion

The Sampo model is useful from the end-users' view point in several ways:

1. *Global view to heterogeneous, distributed contents*. The contents of different content providers can be accessed through one homogeneous data service.
2. *Automatic content aggregation*. For example, when looking for data about an artist, relevant information may be provided by museum collections, libraries, archives, authority records, ontologies, and other sources.
3. *Semantic search*. Semantic content makes it possible to provide the end-user with more "intelligent" services based on ontologies, such as semantic search, semantic autocompletion, and faceted search.
4. *Semantic browsing and recommendations*. Semantic content and network also facilitates semantic browsing and recommendations for additional information.
5. *Other intelligent services*. Also other kind of intelligent services can be created based on machine interpretable content, such as knowledge and association discovery, personalization, and semantic visualizations.

Semantic collaborative portals are very attractive from the content publishers' viewpoint, too:

1. *Distributed content creation.* Semantic technologies can be used for harvesting and aggregating distributed heterogeneous content into global content repositories.
2. *Automated link maintenance.* In semantic portals, links can be created and maintained automatically based on the metadata and ontologies.
3. *Shared content publication channel.* A semantic portal can provide the participating organizations with a shared, cost-effective publication channel.
4. *Enriching each other's contents semantically.* Interlinking content between collaborating organizations enriches the contents of everybody "for free".
5. *Reusing aggregated content.* The content aggregated into a semantic data service can be reused in different applications and cross-portal systems.

However, obtaining interoperability requires more disciplined use of standards, harmonized metadata models, shared vocabularies, and shared best practices. Furthermore, when aggregating content across organization boundaries more collaboration and harmonization of data is needed. The main challenge is often rather organizational than technical: changing, e.g., cataloging practices is not easy, and if changes are made, there is the question of what to do with already cataloged legacy metadata. A practical difficulty is that content management systems in use do not support creation of semantic web data.

Developing more intelligent applications also sets new challenges: such systems are typically more complex, require specific skills from programmers, new tools, and are typically also computationally more complex and not necessarily scale up so easily.

Data enrichment via linked data is promising, but in practice the datasets available have quality problems: many of them, such as DBpedia, have been produced automatically by machines without human touch. A problem here is that the URI identifiers used for concepts (e.g. persons and places) in different datasets are typically different, and the data mappings are not complete or contain errors. Domain ontologies used may be incomplete, violate semantic constraints of standards, and so on. Finally, when reusing content as services one becomes dependent of an external system controlled by somebody else, and often have to be content with suboptimal API services, quality of service, licensing policies, etc.

However, it is clear that these challenges need to be tackled in one way or another when integrating collection data on a semantic level. Semantic web technologies provide a standard approach and a tool set that has already been successfully applied for the task, so why not try it instead of starting from scratch and possibly ending up reinventing similar solutions again.

References

1. Crymble, A., Gibbs, F., Hegel, A., McDaniel, C., Milligan, I., Posner, M., (Eds), W.J.T.: The Programming Historian. 2nd ed (2015), http://programminghistorian.org/
2. Dominque, J., Fensel, D., Hendler, J.A. (eds.): Handbook of Semantic Web. Springer–Verlag (2011)
3. Graham, S., Milligan, I., Weingart, S.: Exploring big historical data. The historian's macroscope. Imperial College Press (2015)
4. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Morgan & Claypool, Palo Alto, CA, USA (2011), http://linkeddatabook.com/editions/1.0/
5. Hyvönen, E., Mäkela, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S.: MuseumFinland—Finnish museums on the semantic web. Journal of Web Semantics 3(2), 224–241 (2005)
6. 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 5th European Semantic Web Conference (ESWC 2008). Springer–Verlag (2008)
7. Hyvönen, E.: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Morgan & Claypool, Palo Alto, CA, USA (2012)
8. 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, Proceedings. Archives and Museum Informatics, Toronto (2009)
9. Mäkelä, E., Hypén, K., Hyvönen, E.: BookSampo—lessons learned in creating a semantic portal for fiction literature. In: Proceedings of ISWC-2011, Bonn, Germany. Springer–Verlag (2011)
10. 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) (2012)
11. Saarti, J.: Aspects of Fictional Literature Content Description: Consistency of the Abstracts and Subject Indexing of Novels by Public Library Professionals and Client (in Finnish). Ph.D. thesis, University of Oulu, Finland (1999)
12. Serola, S., Vakkari, P.: Yleinen kirjasto kuntalaisten toimissa; Tutkimus kirjastojen hyödyistä kuntalaisten arkielämässä. Finnish Ministry of Education and Culture (2011)