CULTURESAMPO
–A Collective Memory of Finnish Cultural Heritage on the Semantic Web 2.0

Eero Hyvönen, Eetu Mäkelä, Tomi Kauppinen, Olli Alm, Jussi Kurki,
Tuukka Ruotsalo, Katri Seppälä, Joeli Takala, Kimmo Puputti, Heini Kuittinen,
Kim Viljanen, Jouni Tuominen, Tuomas Palonen, Matias Frosterus,
Reetta Sinkkilä, Panu Paakkarinen, Joonas Laitio, Katariina Nyberg

Semantic Computing Research Group (SeCo),
Helsinki University of Technology (TKK) and University of Helsinki
`first.last@tkk.fi, http://www.seco.tkk.fi/`

Abstract. This paper presents the Semantic Web 2.0 application CULTURESAMPO, an ambitious system of creating a collective semantic memory of the cultural heritage of a nation on the Semantic Web 2.0, combining ideas underlying the Semantic Web and the Web 2.0. The system addresses the semantic challenge of aggregating highly heterogeneous, cross-domain cultural heritage into a semantically rich intelligent system for human and machine users. At the same time, CULTURESAMPO is an approach to solve the social and practical Web 2.0 challenge of organizing the underlying collaborative ontology development and content creation work of memory organizations and citizens.

1 Components of a National Semantic Memory

In our view, a cultural heritage memory on the Semantic Web 2.0 should be built on three pillars: First we need a cross-domain content infrastructure of ontologies, metadata standards, and related services, that is developed and maintained on a global level through collaborative local efforts. Second, the process of producing ontologically harmonized metadata should be organized in a collaborative fashion, where distributed content producers are able to create semantically correct annotations cost-efficiently through centralized services. Third, the contents should be made available to human end-users and machines thought intelligent search, browsing, and visualization techniques in a portal. For machines, easy to use mash-up APIs and web services should be available. In this way, the collaboratively aggregated, semantically enriched national memory can be exposed and reused easily as services in other portals and applications in the same vein as Google Ads or Maps[1].

CULTURESAMPO[2] [1], released on the public web on September 25, 2008, is an operational demonstration on a national Finnish level of implementing of such a semantic collective memory. The system consists of three components corresponding to the three pillars above:

---

[1] http://code.google.com/apis/maps/

[2] http://www.kulttuurisampo.fi/

1. **National cross-domain content infrastructure FinnONTO** The basis of CULTURESAMPO is the national FinnONTO infrastructure [2, 3] that includes a collaboratively created system of cross-domain ontologies and related ontology services for utilizing them cost-efficiently [4]. The ontologies and the services were released in several languages on the public web as the National Ontology Service ONKI[3] on September 12, 2008.

2. **A content creation process** Our model consists of a set of metadata models and a content creation process for producing and harvesting content from museums, libraries, archives and other organizations, as well as from individual citizens and Web 2.0 sources, such as Wikipedia[4] and Panoramio[5].

3. **Semantic Web 2.0 portal CULTURESAMPO** The portal itself is unique in its use of versatile cross-domain semantic models, new semantic searching and browsing methods, and semantic visualizations for the end-users, for both humans and machines.

In the following these three components are explained in more detail.

2   A Collaborative Ontology Infrastructure

An integral part of CULTURESAMPO are the ontologies and services of the FinnONTO infrastructure [2, 3]. The general idea of the FinnONTO approach is to extend the generic, logic based W3C recommendations with domain specific ontologies in different domains. Most of the FinnONTO ontologies were developed by transforming nationally used thesauri into light weight ontologies. The process was not mechanical like e.g. in [5], but manual processing was required in order to refine the semantic thesaurus relations into full blown subsumption hierarchies.

In the FinnONTO model, the ontologies are developed in a distributed fashion by collaborating expert groups of different fields, and are mapped together to form a large national ontology called KOKO encompassing all domains. At the moment, KOKO includes an upper ontology YSO (20 600 concepts), a museum ontology MAO (6800 concepts), an agriforestry ontology AFO (5500 concepts), an applied art ontology TAO (2600 concepts) and a photography ontology VALO (1900 concepts). The ontologies are provided to end-users not only in the RDF/OWL form, as usual, but as ready to use semantic web widgets [6] using Web 2.0 AJAX APIs, and through conventional Web Services [4]. The KOKO ontology as a single whole is used by developers and end-users, and is the ontological basis of CULTURESAMPO.

In addition to KOKO, CULTURESAMPO also utilizes a geographical registry of 800 000 places in Finland, a spatio-temporal ontology of Finnish counties 1865–2007 [7], an ontology of persons and organizations, and ontologized international systems such as the Iconclass[6] and the Union List of Artists Names (ULAN)[7].

---

[3] http://www.yso.fi/

[4] http://www.wikipedia.org/

[5] http://www.panoramio.com/

[6] http://www.iconclass.nl/

[7] http://www.getty.edu/research/conducting_research/vocabularies/ulan/

3 Content Creation Process

CULTURESAMPO contains cultural objects of 26 different content types: artifacts, paintings, drawings, sculpture, abstract art, novels, comics, web pages, three types of folklore, five types of folk music, photos, aerial photos, persons, organizations, biographies, historical events, skills, videos, buildings, and archeological sites. These content types are represented using 16 different metadata schemas. The aggregated knowledge base contains (Sept 26, 2008) 52 267 cultural objects and 234 597 other resources, such as ontological class concepts and place instances. The cultural objects are described by 624 021 property triples. The content is enriched using reasoning, resulting in 5 844 153 property triples. The enriched knowledge base is used for intelligent information retrieval and for creating semantic recommendation links between objects. In addition, there are 1 194 230 pieces of data related to reference resources. The content is represented using RDF and OWL, and SPARQL is used for recommendations. The system also utilizes external web resources: all Wikipedia articles (in English and Finnish) that have coordinate information, as well as photographs from the Panoramio service can be found on CULTURESAMPO's map views.

These information sources have diverse ownerships. The contents come from 21 museums, archives, and libraries, most of which produce their contents independently from each other using heterogeneous cataloging systems and practices, e.g. different vocabularies. Wikipedia and Panoramio content is created internationally by the public. CULTURESAMPO also has an internal commenting facility by which individuals can contribute new knowledge to individual content items, e.g. identify persons in an old photograph of a museum collection. In these ways, citizens are able to contribute to the national semantic memory. Furthermore, interactive content production based on the SAHA editor [8] is being implemented in the system—this content creation channel has already been used internally in the system by participating organizations.

From a semantic modeling viewpoint, a research focus of our work has been event- and process-based annotations used in artificial intelligence and knowledge representation [9]. In our case, events have been used for modeling cultural processes and narrative stories [1, 10] and for metadata schema integration [11, 12]. The KOKO ontology was designed to support this by clearly separating events and processes from the other concepts along the model of Dolce [13]. In some metadata schemas of CULTURESAMPO it is possible to annotate content using processes in terms of events, subevents and their sequences; the model in use in the prototype is a simplified version of our earlier model [10]. The portal then automatically generates an interactive representation of the process as a kind of a temporal table of contents. This system is used in the prototype for creating skill descriptions, cultural process description, and documentation of processes in videos:

1. **Semantic skill models.** An example of a skill model is the model "Production of Ceramics" produced by experts at the University of Applied Arts in Helsinki. It illustrates and explains the composition of different work phases when manufacturing ceramics. At each phase, semantic recommendations to relevant CULTURESAMPO contents can be created. For example, links to products in collections that were manufactured using the same techniques, are automatically obtained.

2. **Semantic models of cultural processes.** There is a similar kind of chronological model "A Year on a Farm" of the seasonal events and processes taking place at a typical farm in Finland. Again, tools and other materials from CULTURESAMPO are linked automatically with related events. This model was created by a farmer employed at the Finnish Museum of Agriculture. The exhibition of this museum is actually organized using the same idea of presenting farming events taking place during different yearly seasons.

3. **Semantic documentation on videos.** The annotation model can be applied also to documenting *instances* of actual skill events or processes documented on a video. The case example available on the portal describes how the shoemaker Onni Wirlander manufactured a pair of traditional leather boots. The video was produced, and the actual annotation created, by the Espoo City Museum using the SAHA editor connected to the ONKI ontology services. The semantic model describes what happens at different (sub)sequences on the video. Semantic search can find not only the video as a black box, as in systems such as YouTube, but also points of interest *inside* the video. The video can be viewed directly starting from different points of interest. This is important with longer videos. The Wirlander video e.g. lasts over 20 minutes. When watching the video, the recommendation system creates dynamically, for each subsequence separately, recommendation links to materials of interest in the portal, such as tools related to the sequence.

The examples above demonstrate our new idea of representing and storing *immaterial, procedural cultural heritage* in the memory system, here descriptions of handicraft skills and cultural processes. Typical cultural heritage portals contain metadata only about concrete objects.

Content creation in CULTURESAMPO, both ontologies and the metadata, is based on distributing the work to organizations and citizens in a Web 2.0 fashion. In this strategy, extra costs can be minimized by reorganizing the work done already in the organizations and in public. The work is supported by a number of generic FinnONTO tools, such as the metadata editor SAHA, information extraction tool POKA[8], and the semantic content validator VERA[9]. The contents are published with a most liberal open source licensing policy where only the mentioning of the source developers is required. Several organizations that have been developing traditional thesauri are now consedering to start developing light weight ontologies within the FinnONTO framework.

4 The Semantic Web 2.0 Portal CULTURESAMPO

The portal is an end-user application for both 1) human users and 2) machines:

Human User Interface Human clients include the public and researchers, who can use the system for not only finding objects but also for answering questions and analyzing the results along ontological classifications. The system is multilingual (Finnish, Swedish, and English). However, because most of the ontologies and contents are available only in Finnish, the system is not equally powerful in other languages.

---

[8] http://www.seco.tkk.fi/tools/poka/

[9] http://www.seco.tkk.fi/services/vera/

In the human interface, there is always available an easy to use Google-like search input field for typing in any query. However, semantic autocompletion [14] is used, semantic query expansion and reasoning is employed during search, and the underlying ontologies are used to classify and organize the results into meaningful categories. For example, by searching with a persons name, the results are categorized by the roles connecting the person and matched objects, e.g., paintings created by her, sculpture depicting her (but created by someone else), biography or Wikipedia page telling about her etc.

A major novelty that the portal provides is the access to contents through nine "thematic perspectives" available at the main entry page and as choices in the menu bar:

Map views. There are 4 map views available using Google Maps [7]: One for viewing all objects and filtering them in terms of their relation to the places; one for finding old Finnish counties with digitized limits; one for viewing historical maps layered semi-transparently over modern Google maps ; one for finding nearby objects of interest. It is also possible to find Wikipedia articles and Panoramio photos by using the map interfaces.

Relational search. This view is a demonstration of relational search [15], where the idea is not to search for objects but associative relation chains between objects. We used the ULAN registry of 120 000 artists and organizations with 390 000 names. The user types in two names, using semantic autocompletion, and CULTURESAMPO tells how the persons or organizations are related with each other by the social network based on some 50 different social roles (e.g., parent-of, teacher-of, patron-of etc.). The underlying social RDF/OWL network can also be browsed by a graphical network browser.

Domain-centric faceted search. In this view, a new kind of generalized version of the faceted-search paradigm has been developed, called domain centric search [16]. In this paradigm, domain ontologies and the myriad of different properties and roles related to the 16 metadata schemas in use, can be used for very flexible queries and for analysis of contents based on classifying the search results. This advanced search facility is useful e.g. in analyzing the contents. For example, it easy to find our what themes were most popular in the Finnish novels publishes in 2007, or what colors were used in different textile types during different time periods.

Collections view. Here the contents can be accessed based on an organizational view. Each participating organization has an automatically generated home page in the system with links to subcollections and the actual collection items.

Finnish history view. This view is based on an ontology representing events in the Finnish history. These events are of interest of their own, but are also used to create semantic recommendations to other CULTURESAMPO contents, e.g. to biographies of persons participating in the events.

Skills and processes. This view is used for finding cultural procedural descriptions in the system, i.e. semantic models of skills, processes, and documentations on videos. In the prototype, there is one example of each type available.

Biographies. In this view biographies of the National Biography are used to access CULTURESAMPO contents. When reading biographies, related contents are shown based on the concepts extracted from the text using the information extraction tool POKA.

Semantic Kalevala. This view [17] contains a semantically annotated version of the national epic of Finland, Kalevala, that is related in many ways to Finnish art

and culture. The epic also has interesting links to old Finnish folklore, on which it is actually based, and to folk music lyrics. Thousands of runes and pieces of folk music are available in the portal. When reading Kalevala, annotations related to its subsequences can be viewed to help reading, and semantic recommendations to related materials in CULTURESAMPO are automatically produced.

Karelia view. This view contains Wikipedia articles about the Karelia area in Finland that has been influential to Finnish culture. For example, lots of folklore has been collected from this area, and the Kalevala epic is strongly associated there. Like in the biographies view, the POKA system is used for extracting concepts from the texts (here web pages) and for generating related semantic recommendations for more information.

Machines as End-users In addition to human end-users, the system can be used by machines via AJAX interface. The idea is that collaboratively aggregated and semantically enriched national knowledge base can be used not only by the CULTURESAMPO portal but also by other portals and systems on the web. The key idea here is to use cost-effective, ready to use Web 2.0 mash-up services in the same spirit as Goolge Ads or Maps are used on external web pages and applications. In this way, museums, libraries, tourism portals, news papers, individual citizens, and other users can include CULTURESAMPO materials, such as semantic search results and recommendation links, on their web pages using mash-ups. This is a clear win-win situation to everybody: the materials of the CULTURE-SAMPO collaborative network get more visibility and the external users can enrich their own materials for "free", only on 1-2 lines of Javascript code is needed. We envision, that in the near future applications of CULTURESAMPO through its AJAX-interface will be available not only in other cultural heritage portals, but in commercial systems such as tourism and newspaper portals, too. In our own work GPS based mobile and navigation services are being developed, e.g., as a part of the SmartMuseum[10] EU-project.

5   Discussion and Conclusions

The vision and implementation of CULTURESAMPO goes beyond current semantic web portals for cultural heritage [18], rewarded e.g. at the Semantic Web Challenge competition before (MuseumFinland 2004 [19], MultimediaN 2006 [20], and CHIP Demonstrator 2007 [21]. The novelty of the CULTURESAMPO system has many facets: 1) It is highly cross-domain with 26 content types and 16 metadata schemas (usually only one schema such as Dublin Core or VRA is used), 2) it makes use of sophisticated semantic annotation models including events and processes, 3) it uses new kind of semantic search and recommendation techniques, 4) it has exceptionally versatile selection of semantic visualizations available (different map views, timelines, graphs, process visualization, semantic video viewing) 5) it is based on a large nation wide collaboratively maintained infrastructure of ontologies and ontology services, 6) it includes a model of and tools for collaborative semantic content creation, and 7) the services are available for machines, too.

---

[10] http://www.smartmuseum.eu/

Two user evaluation studies concerning semantic recommendations of an earlier version of CULTURESAMPO have been performed with some promising results [11]. Formal evaluation of the now finished prototype with its various new features can now be started.

The portal is scalable in terms of the Web 2.0 content creation model, different types of content, and in terms computational complexity and the number of simultaneous users. The implementation is based on conventional search engine technology (Apache Lucene) that does semantic search using semantic indexing.

References

1. Hyvönen, E., Ruotsalo, T., Häggström, T., Salminen, M., Junnila, M., Virkkilä, M., Haaramo, M., Kauppinen, T., Mäkelä, E., Viljanen, K.: CultureSampo—Finnish culture on the semantic web. The vision and first results. In: Klaus Robering (Ed.), Information Technology for the Virtual Museum. LIT Verlag. (2008)

2. 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, Springer–Verlag (2008)

3. Hyvönen, E.: Semantic interoperability of content and services. A national cross-domain content infrastructure for the semantic web. In: Semantic Computing, IEEE Press / Wiley, New York (2008, forth-coming)

4. Viljanen, K., Tuominen, J., Känsälä, T., Hyvönen, E.: Distributed semantic content creation and publication for cultural heritage legacy systems. In: Proceedings of the 2008 IEEE International Conference on Distibuted Human-Machine Systems, Athens, Greece, IEEE Press (2008)

5. van Assem, M., Malaise, V., Miles, A., Schreiber, G.: A method to convert thesauri to SKOS. In: Proceedings of the Third European Semantic Web Conference (ESWC'06). Lecture Notes in Computer Science (2006)

6. Mäkelä, E., Viljanen, K., Alm, O., Tuominen, J., Valkeapää, O., Kauppinen, T., Kurki, J., Sinkkilä, R., Känsälä, T., Lindroos, R., Suominen, O., Ruotsalo, T., Hyvönen, E. In Nixon, L., Cuel, R., Bergamini, C., eds.: First Industrial Results of Semantic Technologies, proceedings, co-located with ISWC 2007 + ASWC 2007, Busan, Korea. (2007) CEUR Workshop Proceedings, Vol 293.

7. Kauppinen, T., Väätänen, J., Hyvönen, E.: Creating and using geospatial ontology time series in a semantic cultural heritage portal. In: Proceedings of the ESWC 2008, Tenerife, Spain, Springer–Verlag (2008)

---

[11] http://www.seco.tkk.fi/projects/finnonto/

8. Valkeapää, O., Alm, O., Hyvönen, E.: Efficient content creation on the semantic web using metadata schemas with domain ontology services. In: Proceedings of ESWC 2007, Innsbruck, Austria, Springer–Verlag (2007)

9. Sowa, J.: Knowledge Representation. Logical, Philosophical, and Computational Foundations. Brooks/Cole (2000)

10. Junnila, M., Hyvönen, E., Salminen, M.: Describing and linking cultural semantic content by using situations and actions. In: Klaus Robering (Ed.), Information Technology for the Virtual Museum. LIT Verlag. (2008)

11. Ruotsalo, T., Hyvönen, E.: A method for determining ontology-based semantic relevance. In: Proceedings of the International Conference on Database and Expert Systems Applications DEXA 2007, Regensburg, Germany, Springer (2007)

12. Ruotsalo, T., Hyvönen, E.: An event-based approach for semantic metadata interoperability. In: Proceedings of the ISWC 2007 + ASWC 2007, Busan, Korea, Springer–Verlag (2007)

13. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening ontologies with DOLCE. In: Proc. of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, Springer–Verlag (2002)

14. Hyvönen, E., Mäkelä, E.: Semantic autocompletion. In: Proceedings of the first Asia Semantic Web Conference (ASWC 2006), Beijing, Springer-Verlag (2006)

15. Kurki, J., Hyvönen: Relational semantic search: Searching social paths on the semantic web. In: Poster Proceedings of the ISWC + ASWC 2007, Busan, Korea. (2007)

16. Mäkelä, E., Suominen, O., Hyvönen, E.: Automatic exhibition generation based on semantic cultural content. In: Proc. of the Cultural Heritage on the Semantic Web Workshop at ISWC + ASWC 2007. (2007)

17. Hyvönen, E., Takala, J., Alm, O., Ruotsalo, T., Mäkelä, E.: Semantic Kalevala—accessing cultural contents through semantically annotated stories. In: Proceedings of the Cultural Heritage on the Semantic Web Workshop at the 6th International Semantic Web Conference (ISWC 2007), Busan, Korea. (2007)

18. Hyvönen, E.: Semantic portals for cultural heritage. In: Handbook of Ontologies, 2. edition, Springer-Verlag (2008, forth-coming) Chapter under review.

19. 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) (2005) 224–241

20. Schreiber, G., Amin, A., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Hollink, L., Huang, Z., van Kersen, J., de Niet, M., Omelayenko, B., van Ossenbruggen, J., Siebes, R., Taekema, J., Wielemaker, J., Wielinga, B.J.: MultimediaN e-culture demonstrator. In: The Semantic Web - Proceedings of the 5th International Semantic Web Conference 2006. (2006) 951–958

21. Wang, Y., Stash, N., Aroyo, L., Gorgels, P., Rutledge, L., Schreiber, G.: Recommendations based on semantically-enriched museum collection. Journal of Web Semantics (2008) (in press)