HealthFinland—Publishing Health Promotion Information on the Semantic Web

Eero Hyvönen[1], Kim Viljanen[1], Osma Suominen[1], and Eija Hukka[2]

[1]Semantic Computing Research Group (SeCo)
Helsinki University of Technology (TKK) and University of Helsinki
[2] National Public Health Institute, Finland

Abstract

This paper shows how semantic web techniques can be applied to solving problems of distributed content creation, discovery, linking, aggregation, and reuse in health information portals, both from end-user's and content producer's viewpoints. As a case study, the national semantic health portal HealthFinland is presented. It provides citizens with intelligent searching and browsing services to reliable and up-to-date health information created by various expert organizations and authorities in the field of health promotion in Finland. The system is based on a shared semantic metadata schema, ontologies, and mash-up ontology services. The content includes the metadata of thousands of web documents such as web pages, articles, reports, campaign information, news, services, and other information related to health.

1. Introduction

Health information on the web is provided by different organizations of varying levels of trustworthiness, is targeted to both laymen and experts, is available in various forms, and is written in different languages. The difficulty of finding relevant and trustworthy information in this kind of heterogeneous environment creates an obstacle for citizens concerned about their health.

This paper discusses problems concerning semantic information portals [1] when publishing health information on the web for the citizens. We consider both the publishers' and the end-users' viewpoints. A distributed semantic web[1] content publishing model has been developed for health promotion organizations, based on a shared metadata schema, ontologies [2], and mash-up ontology services, by which the content is cost-effectively created by independent content producers at different locations. Our system aggregates and makes the content semantically interoperable, as in [3], to be reused in different applications without modifying it.

To test and demonstrate the approach, we have created an operational prototype of the national semantic health information portal HealthFinland—Finnish Health Information on the Semantic Web[2] (HF) [14]. The content for the prototype (ca. 6000 web documents) was created by the National Public Health Institute (KTL)[3], the UKK Institute for Health Promotion Research[4], the Finnish Institute of Occupational Health[5], the national Suomi.fi citizen's portal[6], and the Ministry of Justice[7]. More organizations are joining in the system.

2. A New Semantic Web Approach

In traditional web publishing, content creators publish web pages and link them together independently from each other. Content management systems (CMS) and portals are utilized to aggregate related content within one site, and to provide local search and linking services. Linking between sites is usually done manually. Search engines are utilized to provide content aggregation services on the global cross-site level. In HF, we wanted to create a new kind of collaborative distributed content creation model for publishing health related information on the web.

Firstly, we try to minimize duplicate redundant work and costs in creating health content on the national level by producing it only once by one organization, and by making it possible to re-use the content in different web applications by the other organizations, not only in the organization's own portal. This possibility is facilitated by annotating the content locally with semantic metadata based on shared ontologies, and by making the global repository available for other applications, such as HF. This is a generalization of the idea of "multi-channel publication" of XML, where a syntactic XML structure can be rendered in different ways. In our case, semantic RDF content is re-used through *multi-application*

[1] http://www.w3.org/2001/SW/

[2] http://www.seco.tkk.fi/applications/tervesuomi/

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

[4] http://www.ukkinstituutti.fi/

[5] http://www.ttl.fi/

[6] http://www.suomi.fi/

[7] http://www.finlex.fi/

publication. A promising cost-effective way of re-using content as ready-to-use interface components is mash-up web services [15].

Our second goal is to minimize the maintenance costs of portals by letting the computer take care of semantic link maintenance and aggregation of content from the different publishers. This possibility is also based on shared semantic metadata and ontologies. New content relevant to a topic may be published at any moment by any of the content providers, and the system should be able to put the new piece of information in the right context in the portal, and automatically link it with related information.

The third major idea of HF is to provide the end-user with intelligent services for finding the right information based on his or her own conceptual view to health matters, and for browsing the contents based on their semantic relations. The views and vocabularies used in the end-user interface may be independent of the content providers' organizational perspective, and are based on layman's vocabulary that is different from the medical expert vocabularies used by the content provider and information specialist in indexing the content.

In the HF system, the content providers produce web pages, documents, and other resources along their organizational interests and for their own purposes. However, the content is annotated by using a shared metadata schema and ontologies for the others to use, too. Selected content is then harvested into a global RDF knowledge base to be re-used in other portals. In this paper, we consider one portal in particular: the semantic portal HF that provides citizens with information services based on the global health information repository.

3. Metadata and Ontologies

The ontological infrastructure of HF consists of two major components: 1) A metadata schema [5] that specifies what elements are used for describing the web documents to be included in the system, and what kind of values the elements (properties) can take. The metadata schema is shared by all organizations creating the content and ensures the *syntactic interoperability* of the content. 2) A set of ontological vocabularies whose concepts are used to fill in the values of the metadata schema. The ontologies are also shared by the organizations and their usage ensures the *semantic interoperability* of the content.

The metadata schema of HF is based on the Dublin Core Element Set[8], along with refinements introduced in DCMI Terms[9]. In addition, to allow a more detailed description of the required metadata, we have introduced three extensions to Dublin Core elements: *ts:genre* field (namespace *ts* refers to HF) for the content genre, *ts:url* for document locators, and *ts:isTranslationOf* for presenting the relation between language translations of documents. The metadata is presented using RDF[10]. A subset of the metadata can also be embedded in (X)HTML pages using META and LINK elements based on the Dublin Core recommendation[11].

The metadata is published by making it available on a public WWW server where it can be collected by the HF metadata harvester into a centralized metadata server. During the harvesting, 1) the content is transformed into RDF (if originally presented in HTML), 2) missing values are replaced with default values when possible, and 3) the RDF is validated against the metadata schema and other validation rules. Each metadata producer gets a report of warnings, errors, and other problems if encountered during harvesting and validating the content. If some parts or all of the metadata is unacceptable due to serious errors, the metadata is discarded until necessary corrections are made.

Semantic interoperability in HF is obtained by using a set of shared ontologies for filling in the values of the metadata schema. The ontologies include a Medium Ontology containing resources for representing different media types (only web pages are considered at the moment), an Audience Ontology representing categories of people, such as gender, professional groups, risk groups, and age groups, a Place Ontology containing geographical places (e.g., Finland, Helsinki) in a part-of hierarchy, a Genre Ontology for genre types (e.g., news, game), DCMI type ontology for media types (e.g., text, sound, video), and a Time Ontology. The most important ontologies in HF are the three core subject domain ontologies that are used for describing the subject matter of web contents:

1) The Finnish General Upper Ontology (YSO)[12] [7] that includes approximately 20,000 concepts based on the General Finnish Thesaurus YSA[13].

2) The international Medical Subject Headings (MeSH)[14] vocabulary of 23,000 concepts, with a Finnish translation, FinMeSH, acquired from the Finnish Medical Society Duodecim[15.] This vocabulary was transformed into the SKOS Core[16] format without changing the semantics of the vocabulary or its structure.

3) The European Multilingual Thesaurus on Health

[8] http://dublincore.org/documents/dces/

[9] http://dublincore.org/documents/dcmi-terms/

[10] http://dublincore.org/documents/dcq-rdf-xml/

[11] http://dublincore.org/documents/dcq-html/

[12] http://www.seco.tkk.fi/ontologies/yso/

[13] http://www.vesa.lib.helsinki.fi/

[14] http://www.nlm.nih.gov/mesh/

[15] http://www.duodecim.fi/

[16] http://www.w3.org/2004/02/skos/core/

Promotion (HPMULTI)[17], which included a Finnish translation. HPMULTI contains approximately 1200 concepts related specifically to health promotion. Also HPMULTI was transformed into SKOS/RDF format.

All three ontologies were needed to cover the subject matter of the portal properly. YSO is broad but too general w.r.t. detailed medical content. On the other hand, MeSH contains lots of useful medical concepts, is widely used in the health sector, but is focused on medical terminology. HPMULTI complements the two vocabularies by focusing on health promotion terminology.

3. Content Creation

In HF three ways of creating metadata are supported:

1) The content can in many cases be transformed fairly accurately into ontological form automatically from a content provider's CMS. This is because the HF metadata schema is strongly based on Dublin Core and because many content providers in Finland use thesauri (e.g., YSA and MeSH). For example, some legal content produced by the Finnish Ministry of Justice is harvested for HF in this way.

2) Existing web content management systems (CMS) can be connected cost-effectively to the ontology mash-up services of the ONKI Ontology Server framework[18] [7]. ONKI provides ontological functionalities, such as concept searching, browsing, disambiguation, and fetching, as ready-to-use mash-up components that communicate asynchronously by AJAX[19]. Only a short snippet of JavaScript code need to be added to the web page for exploiting the ONKI services [15].

3) We have created a centralized browser-based annotation editor SAHA [4] for annotating web pages. It is useful for HF content providers who do not have a CMS or cannot add ONKI AJAX mash-up ontology support to their CMS.

3. Semantic Services for End-users

The HF user interface is based on the faceted browsing paradigm [8,9,12,3]. A challenge in publishing health-related information in a citizens' semantic portal is the gap between the citizens' information needs and the professional conceptualizations and terminology used in medical ontologies. To bridge this gap and to enable an intuitive facet-based user interface for the portal, we constructed the search facets by using a card sorting method [10] to find out how users tacitly group and organize concepts in the health domain. The new user-centric facets organize the material from a citizens' point of view, and they are mapped by the portal to concepts in the medical ontologies.

The HF portal provides two basic end-user services:

1) *Semantic search.* A faceted search engine based on the semantics of the content. The main facets of the portal are Topic, Life event, Group of people, and Body part. In addition, secondary drop-down facets for constraining the search with a set of additional choices, are provided for Genre, Publisher, Publication year and Audience. Furthermore, keyword searches can be initiated at any point and can be combined with category browsing. Traditional keyword search functionality has been semantically enhanced by targeting not only content titles, descriptions and body text but also the facet categories and underlying ontology concepts, including non-preferred concept labels. Thus, synonyms and abbreviations can be used in keyword searches provided they are known in the ontology.

2) *Semantic browsing*. Automatic dynamic linking between pages is provided based on the semantic relations in the underlying knowledge base. The portal also provides semantic recommendation links at several stages: 1) individual content items (pages) are linked to related material, 2) search result listings provide "best picks", and 3) concept pages link to related content. Recommendations are generated using ontological knowledge and grouped according to genre (e.g. statistics, research activities, news items, laws) or language (e.g. similar content in English).

The HF portal also incorporates an alphabetical index of concepts as well as a concept browser that can be used to browse the subject ontology and for concept-based search of content.

The portal is implemented as a Java Servlet application running on Apache Tomcat[20]. It is built using the Tapestry framework[21] and uses Jena[22] for RDF functionality. Search and recommendation functionality has been implemented using the Lucene search engine[23], which has been enhanced to handle category and concept queries.

4. Summary

HF addresses the problems finding of health related information on the web as follows: 1) Content finding is supported by cross-portal semantic search, based on concepts and facets rather than keywords. 2) The problem of outdated and missing links is eased by providing the end-user with semantic recommendations that change

[17] http://www.hpmulti.net/

[18] http://www.seco.tkk.fi/services/onki/

[19] http://en.wikipedia.org/wiki/Ajax_(programming)

[20] http://tomcat.apache.org/

[21] http://tapestry.apache.org/

[22] http://jena.sourceforge.net/

[23] http://lucene.apache.org/

dynamically as content is modified. 3) Content aggregation is facilitated by end-user facets that collect distributed but related information from different primary sources. 4) Quality of content is maintained by including only trustworthy organizations as content producers. 5) End-user's expertise level is taken into account by the metadata element “Audience”. Separation of end-user vocabularies from professional indexing vocabularies makes it possible to the citizen search and browse content using layman's vocabulary.

HF also addresses the problems of content providers: 1) Duplication of content production can be minimized by the possibility of aggregating cross-portal content. 2) Reusing the global content repository is feasible by external applications that can reuse the content, such as HF. 3) Internal and external link management problems are eased by the dynamic semantic recommendation system of the portal and the content aggregation mechanisms. 4) The tedious content indexing task is supported cost-effectively by shared ontology mash-up services. 5) Metadata quality can be enhanced by providing indexers with ontology services by which appropriate indexing concepts can be found and correctly entered into the system.

The content creation model presented is based on a shared metadata schema and ontologies as in [3]. However, the idea of sharing ontologies through mash-up ontology services is new. The user interface is based on the faceted search paradigm [8,9,12,3], but integrated with semantic web ontologies and reasoning with semantic recommendations [11], as in [12,3]. A new feature of the system is the separation of end-user facets from indexing ontologies [13,10] , which is crucial in the medical domain. The card sorting approach [10] was found useful in accomplishing this.

This work is a part of the national semantic web ontology project FinnONTO[24] 2003-2007 [7], funded mainly by the National Funding Agency for Technology and Innovation (Tekes) and the Ministry of Social Affairs and Health.

References

[1] Sidoroff, T., Hyvönen, E.: Semantic e-goverment portals - a case study. In: Proceedings of the ISWC-2005 Workshop Semantic Web Case Studies and Best Practices for eBusiness SWCASE05 (2005).

[2] Fensel, D.: Ontologies: Silver bullet for knowledge management and electronic commerce (2nd Edition). Springer-Verlag (2004).

[3] Hyvönen, E., Mäkelä, 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.

[4] Valkeapää, O., Alm, O., Hyvönen, E.: Efficient content creation on the semantic web using metadata schemas with domain ontology services (system description). In: Proceedings of the European Semantic Web Conference ESWC 2007, Innsbruck, Austria, Springer–Verlag (2007).

[5] Suominen, O., Viljanen, K., Hyvönen, E., Holi, M., Lindgren, P.: TerveSuomi.fi:n metatietomäärittely (Metadata schema for TerveSuomi.fi), Ver. 1.0 (26.1.2007). (2007), http://www.seco.tkk.fi/publications/.

[6] Hyvönen, E., Valo, A., Komulainen, V., Seppälä, K., Kauppinen, T., Ruotsalo, T., Salminen, M., Ylisalmi, A.: Finnish national ontologies for the semantic web - towards a content and service infrastructure. In: Proceedings of International Conference on Dublin Core and Metadata Applications (DC 2005) (2005).

[7] Hyvönen, E., Viljanen, K., Mäkelä, E., et al.: Elements of a national semantic web content infrastructure—case Finland on the semantic web. Paper, submitted, http://www.seco.tkk.fi/publications/ (2007).

[8] Pollitt, A.S.: The key role of classification and indexing in view-based searching. Technical report, University of Huddersfield, UK (1998) http://www.ifla.org/IV/ifla63/63polst.pdf.

[9] Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., Lee, K.P.: Finding the flow in web site search. CACM 45(9) (2002) 42–49.

[10] Suominen, O., Viljanen, K., Hyvönen, E.: User-centric faceted search for semantic portals. In: Proc. of ESWC 2007, Innsbruck, Austria, Springer-Verlag, New York (2007).

[11] Viljanen, K., Känsälä, T., Hyvönen, E., Mäkelä, E.: Ontodella - a projection and linking service for semantic web applications. In: Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA 2006), Krakow, Poland, IEEE (2006).

[12] 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, May 10-12, Heraklion, Greece, (forthcoming), Springer–Verlag, Berlin (2004).

[13] Holi, M., Hyvönen, E.: Fuzzy view-based semantic search. In: Proceedings of the 1st Asian Semantic Web Conference (ASWC2006), Beijing, China, Springer-Verlag (2006).

[14] Hyvönen, E., Viljanen, K., Suominen, O.: HealthFinland – Finnish Health Information on the Semantic Webs. Paper, submitted, http://www.seco.tkk.fi/publications/ (2007).

[15] Mäkelä, E., Viljanen, K., Alm, O.: Enabling the semantic web with ready-to-use mash-up components. Paper, submitted, http://www.seco.tkk.fi/publications/ (2007).

Address for correspondence

Eero Hyvönen
Helsinki University of Technology
P.O. Box 5500
FI-02015 TKK, Finland
http://www.seco.tkk.fi/
eero.hyvonen(at)tkk.fi

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