# HealthFinland —A National Publication System for Semantic Health Information

Osma Suominen, Eero Hyvönen, Kim Viljanen, and Eija Hukka (+)

Semantic Computing Research Group (SeCo),
Helsinki University of Technology (TKK), Department of Media Technology
University of Helsinki, Department of Computer Science
firstname.lastname@tkk.fi, http://www.seco.tkk.fi/
(+) National Public Health Institute
firstname.lastname@ktl.fi

**Abstract.** HEALTHFINLAND is a national semantic publishing system for providing Finnish citizens with reliable, up-to-date information about health. The system consists of three parts: 1) a centralized service of health ontologies with tools, 2) a semantic content creation channel based on several distributed health organizations, and 3) an intelligent semantic portal aggegating and presenting the contents from intuitive and health promoting end-user perspectives. The system demontrates how semantic web techniques can be applied to solving problems of distributed content creation, discovery, linking, aggregation, and reuse in health information on a national level, from end-users', content publishers', and machine processing viewpoints. The HEALTHFINLAND prototype is operational on the web, and a production version of the portal, created by the National Health Institute, will be released in February 2009.

## 1 Problems of Mediating Health Information

Health information is one of the most frequently searched material on the web [1]. However, a citizen searching for health information on the web faces many challenges [2]. Health information is often published at organization-centric websites, requiring prior knowledge of the organizations involved. After finding a piece of interesting information, it is often tedious and difficult to find related relevant web resources. Outdated and broken links are common. Satisfying an end-user's information need often requires *aggregation* of content from several information providers. Additionally, the quality and trustworthiness of information varies. In many cases it is difficult know whether a content is based on scientific results or layman opinions and rumors, or whether it is motivated by commercial interests. Research organizations that focus on scientific issues cannot necessarily communicate the results effectively to ordinarily citizens.

From the viewpoint of the health organizations, creating health information to citizens is also problematic in many ways [2]. Several organizations create overlapping content, which wastes time and money and is confusing to the end user. Content in websites is usually minimally annotated for the purpose of presenting it on a particular

site and for the particular purpose of the publisher. This makes it difficult and expensive for other organizations to re-use content across portals even if the portal owners were willing to do this. The problems of maintaining links up-to-date is very costly and tedious, especially when dealing with links to external sites to which the maintainer and the content management system has no control.

If metadata is produced, finding the right keywords and other metadata descriptions for web pages and documents is difficult and time consuming for information producers. The vocabularies used, such as MeSH[1], UMLS[2] or SNOMED CT[3], are very large and require expertise to use.

Furthermore, there are several quality issues involved when publishing health information: 1) Quality of the content creation process (e.g. regular reviews and updates of published material) 2) Quality of the content itself (e.g., errors in the medical subject matter or understandability for the target audience). 3) Quality of additional information on pages (e.g., it is advisable to show the date of publication on each page). 4) Quality of the metadata. For example, one indexer may use only few general keywords while another prefers a longer detailed list, which leads to problems of unbalanced and low quality metadata.

HEALTHFINLAND addresses these problems both from the publishers and citizens viewpoints. Furthermore, the solutions are provided for machines to use through semantic widgets.

## 2 The HEALTHFINLAND Solution

An overview of the HEALTHFINLAND system is given in Figure 1. The system consists of content sources such as health websites, content aggregation infrastructure, ontology services and a citizens' portal. The novelty of the system is based on three major ideas:

First, HEALTHFINLAND minimizes duplicate redundant work and costs in creating health content on a national level by collaboration. A goal of the HEALTHFINLAND collaborative production network is to ensure that information about a health topic is produced only once and by the organization that knows most about it. By using semantic technologies the content can then be re-used in different web portals by the other organizations, not only in the organization's own web site. This possibility is facilitated by annotating the content locally with semantic metadata based on shared ontologies, and by making the global repository available by a semantic portal and as mash-up web services. This is a generalization of the idea of "multi-channel publication" of XML, where a single syntactic structure can be rendered in different ways, but on the semantic metadata level and using RDF: semantic content is re-used through *multi-application publication*.

The second key idea behind HEALTHFINLAND is to try 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

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

[2] http://umlsinfo.nlm.nih.gov

[3] http://www.snomed.org/snomedct/

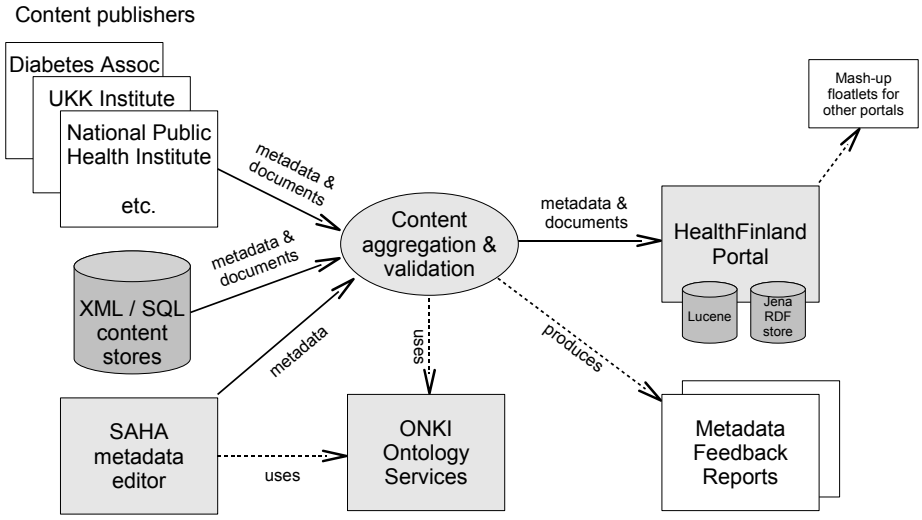


**Figure 1.** HEALTHFINLAND Architecture

published at any moment by any of the content providers, and the system is able 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 HEALTHFINLAND is to provide the end-user with intelligent services for finding the right information based on her own conceptual view to health, and for browsing the contents based on their semantic relations. The views and vocabularies used in the end-user interface are 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 providers in indexing the content.

In the following, the three main components of HEALTHFINLAND—content and service infrastructure, content creation system, and the semantic portal—are shortly explained. The concept and ideas behind HEALTHFINLAND were developed in [2]. The design and implementation of the portal user interface is presented [3], and the design and implementation of the portal in more detail (in Finnish) in [4]. The ontology service infrastructure and services are presented in [5, 6].

## 3 Metadata, Ontology, and Service Infrastructure

The metadata scheme of HEALTHFINLAND is based on the Dublin Core Element Set[4], along with refinements introduced in DCMI Terms[5] and a few domain-specific extensions such as a field for describing content genres. The metadata in HEALTHFINLAND

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

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

is presented using RDF, conforming to the recommendations for expressing Dublin Core in RDF [7, 8]. A subset of the metadata can also be embedded in (X)HTML pages using META and LINK elements based on the Dublin Core recommendation [9].

Semantic interorerability in HEALTHFINLAND is obtained by using a set of shared ontologies for filling in the values of the metadata scheme. The ontologies include a Medium Ontology containing resources for representing different media types (Web page, CD, DVD, etc.), an Audience Ontology representing categories of people, such as sex groups, professional groups, risk groups, and age groups, a Place Ontology containing geographical places (e.g., Finland, Helsinki, etc) in a part-of hierarchy, a Genre Ontology for document types (article, news, campaign etc.), DCMI type ontology media types (text, sound, video etc.), and a Time Ontology. In the future, custom made organizational vocabularies can also be used, provided that they are linked with the HEALTHFINLAND ontologies.

The most important ontologies in HEALTHFINLAND 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)[6] that includes approximately 20 000 concepts. The YSO ontology was created by transforming the General Finnish Thesaurus YSA[7] into RDF/OWL format using the Protégé editor[8] and by manually crafting the concepts into full-blown rdfs:subClassOf hierarchies [6]. 2) The international Medical Subject Headings (MeSH) which includes approximately 23 000 concepts. The vocabulary was transformed into the SKOS Core format[9] without changing the semantics of the vocabulary or its structure using conversion tools and methods developed at the Free University of Amsterdam [10]. 3) The European Multilingual Thesaurus on Health Promotion [10] (HPMULTI), which included a Finnish translation. HPMULTI contains approximately 1200 concepts related specifically to health promotion. HPMULTI was transformed into SKOS/RDF in a similar way as MeSH using a custom conversion tool.

All three vocabularies 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 clinical healthcare. HPMULTI complements the two vocabularies by focusing on health promotion terminology. To make the three vocabularies semantically compatible with each other, the Health Promotion Ontology was built using YSO as the structural basis and extending it manually with concepts from the two other vocabularies. Currently, the Health Promotion Ontology contains all concepts from YSO and HPMULTI together with some 2500 concepts from MeSH.

To enable the usage of ontologies in legacy content management systems, the National Ontology Service ONKI was designed and implemented [11]. A major innovation of ONKI in the Selector Widget by which ontology services can be mashed-up with legacy systems on HTML level as ready-to-use functionalities using AJAX. The idea is related to Google Maps and Ads, but in our case services for concept finding,

[6] http://www.seco.tkk.fi/ontologies/yso/
[7] http://www.vesa.lib.helsinki.fi
[8] http://protege.stanford.edu
[9] http://www.w3.org/2004/02/skos/core/
[10] http://www.hpmulti.net/

semantic disambiguation, and fetching from a centralized service can be used. The National Ontology Service ONKI is available with related services at http://www.yso.fi/. The service now contains lots of other ontologies than those of HEALTHFINLAND, but the system was first piloted as a part of it.

An integral part of the content creation infrastructure is the browser-based metadata editor SAHA [12] that combines a given metadata scheme with ONKI services based on the range restiction on element values. It was used in HEALTHFINLAND for manual annotation of health information.

## 4 Content Creation Process

Eight Finnish health organizations—Finnish Diabetes Association, Duodecim Society, National Public Health Institute, Social and Welfare Services of Oulu, Savonia University of Applied Sciences, Finnish Centre for Health Promotion, Finnish Institute of Occupational Health, and UKK Institute—contibuted to the health contents available in the protype on the web. The system contains almost 3500 health content objects falling into 15 content types: articles, news items, forms, guides, educational materials, organization presentations, projects and campaigns, events, tests, databases, magazine articles, statistics, research, and question-answer pairs.

Depending on the organization and content, the content was created in different ways: two organizations used embedded mark-up conforming to the HEALTHFINLAND metadata scheme on their HTML pages; two organizations provided their content in XML formats that were then transformed into RDF by special transformators; one organization provided a relational database that was similarly transformed into RDF; and three organizations provided the metadata directly in the specified RDF format using the SAHA editor attached to ONKI services. The XML and database sources used different subject vocabularies that had to be mapped to the HEALTHFINLAND subject ontologies before the transformation could be completed.

To support local content creation, a validator service was implemented by which validity of RDF metadata can be checked. This tool finds errors in data and provides a report to the metadata producer for corrections. The content was published in the portal only after the metadata is approved by the validator.

The content creation infrastructure was evaluated by collecting user feedback from annotators after several months of use. Overall, the processes and tools were considered functional, but several improvements were also suggested.

The metadata scheme and related annotation tools could be improved by removing or hiding some optional fields (e.g., Coverage) that were seldom used. Manual work could be reduced using semiautomatic annotation tools to suggest, e.g., document languages and subjects based on text analysis. Also, annotators could avoid repetitive work by using previous annotations as templates for annotating new content.

The use of three different complementary subject vocabularied was somewhat impractical as it forced annotators to look up concepts in each vocabulary. Using the Health Promotion Ontology, developed during the project, for directly annotating resources would help; however, it was still being built when annotation work started. In

addition, a process is needed for adding new necessary concepts into the ontologies during annotation.

The feedback reports were considered a useful tool for metadata quality control, but the details shown were often overwhelming. A way to suppress warnings about acknowledged but unimportant problems, such as systematic syntax errors, was requested.

## 5 HealthFinland Portal

The user interface of the portal is based on the view-based semantic search paradigm [13]. A special problem in HEALTHFINLAND is that the ontologies used for annotating the health content are intended for health information professionals to use, whereas the portal is mainly targeted at the general public. This means that the annotation ontologies and their concept hierarchies cannot be used directly for querying in navigational facets, as in semantic portals such as MuseumFinland[11]. Instead, we have constructed new, citizen-centric facets and mapped these to the underlying ontologies [3]. The primary navigational facets are Topic, Life event, Group of people, Body part, Document type (genre), Publisher and Audience. Keyword search can be freely combined with faceted searches.

In addition to faceted search, the portal provides recommended links based on ontological knowledge (e.g. "smoking is a risk factor for lung cancer") and an alphabetical index of concepts.

The user interface is multilingual and supports Finnish, Swedish and English. Cross-language queries are supported, so that, e.g., an English user interface and English-language categories may be used to perform searches in the Finnish-language content items, demonstrating the potential of multilingual ontologies. However, not all facet categories and ontological concepts have been translated into all three languages.

In addition to serving end users directly, the portal also provides a *floatlet*, i.e., an AJAX-based semantic Web widget that can be incorporated into other portals to display related content items from the HEALTHFINLAND system. This way, the portal contents are exported for machine processing. The implementation is similar to the related museum item display described in an earlier publication [5].

The portal is implemented as a Java Servlet application running on Apache Tomcat. It is built using the Tapestry framework and uses Jena for RDF functionality. Search and recommendation functionality has been implemented using the Lucene search engine, which has been enhanced to handle category and concept queries. The use of Lucene for all search tasks makes the portal implementation very efficient and should allow scaling the system to orders of magnitude larger document collections.

The portal application has been evaluated with a series of user tests. The first user interface mock-ups were presented to potential users of the portal and the test subjects were asked to describe the user interface elements they saw. The mock-ups were refined until subjects were able to understand the purpose and function of the user interface.

The first working implementation of the portal was evaluated with a two-phase usability test. The test subjects represented the target audience of the portal, i.e., ordinary

[11] http://www.museosuomi.fi

citizens with varying backgrounds. A pilot test with two users was used to refine the test setup and the information retrieval tasks, expressed as realistic scenarios involving the test subjects.

In the second phase, six users completed four retrieval tasks using the portal. Each task was successfully completed by at least three users, with an average task success rate of 70%. Typical task completion times were 3–10 minutes depending on the difficulty of the task. The vast majority of the first actions of test subjects took them closer to their goal, indicating that the portal front page design was succesful. However, deeper in the topic hierarchy the subdivision of topics was often confusing or overwhelming, as it reflected the structure of the underlying ontology. Most users did not make use of secondary query facets or recommendation links between documents, commenting that they didn't seem to help them achieve their task goals. The alphabetical index was seldom used, but proved useful for the users who tried it. A post-test questionnaire of subjective usability was used to calculate a System Usability Scale [14] score of 70,8 points, indicating that users generally liked using the portal.

## 6 Discussion and conclusions

The HEALTHFINLAND system applies state-of-the art semantic web technology into the problems of publishing, aggregating and finding health information on a national level. The system demonstrates how shared meanings, expressed using ontologies, can be used to bind together syntactically and semantically heterogeneous content sources from different publishers. The content creation, validation and aggregation infrastructure, including ontology services and tools, enables the collaborative publication of health content and reduces duplicate work. From an end-user perspective, the underlying semantic technology enables an innovative citizen-centric faceted search user interface as well as more traditional services such as keyword search and topical index. All the tools and services discussed above have been implemented and tested in a real world setting, and the prototype portal has been published on the Web[12]. Further instructions are provided at the project homepage[13].

## 7 Future Work

The National Health Institute is building a production version of the HEALTHFINLAND portal, to be published in February 2009. The new version of the portal utilizes the same content creation infrastructure discussed above, which has been integrated into the Alfresco content management system. The user interface is implemented using Java portlet technology and built using Liferay portal software.

Our ongoing research is focused on building personalization services, such as push e-mail and recommendation links based on user profile information, into the HEALTHFINLAND portal. The portal will also be expanded with a registry of health services that will be semantically interlinked with information resources.

[12] http://demo.seco.tkk.fi/tervesuomi/

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

## Acknowledgements

This work is a part of the national semantic web ontology project FinnONTO[14] 2003-2007, 2008-2010, funded mainly by the National Funding Agency for Technology Innovation (Tekes) and the Ministry of Social Affairs and Health. The HEALTHFINLAND project is co-ordinated by the National Health Institute in Finland (Project Coordinator Eija Hukka). We thank Markus Holi, Petri Lindgren, and Johanna Eerola for their input to the work reported in this paper.

## References

1. Pew Internet & American Life Project: Online Health Search (2006)
2. Hyvönen, E., Viljanen, K., Suominen, O.: HealthFinland—Finnish health information on the semantic web. In: Proceedings of ISWC 2007 + ASWC 2007, Busan, Korea, Springer–Verlag, Berlin (2007)
3. Suominen, O., Viljanen, K., Hyvönen, E.: User-centric faceted search for semantic portals. In: Proceedings of the ESWC 2007, Innsbruck, Austria, Springer–Verlag, Berlin (2007)
4. Suominen, O.: Käyttäjäkeskeinen moninäkymähaku semanttisessa portaalissa (user-centric faceted search in a semantic portal). Master's thesis, University of Helsinki, Department of Computer Science (February 2008)
5. 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)
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 ESWC 2008, Tenerife, Spain, Springer–Verlag, Berlin (2008)
7. Dublin Core Workgroup: Expressing simple Dublin Core in RDF/XML (2002) http://dublincore.org/documents/dcmes-xml/.
8. Dublin Core Workgroup: Expressing qualified Dublin Core in RDF/XML (2002) http://dublincore.org/documents/dcq-rdf-xml/.
9. Dublin Core Workgroup: Expressing Dublin Core in HTML/XHTML meta and link elements (2003) http://dublincore.org/documents/dcq-html/.
10. 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 (June 2006)
11. Viljanen, K., Tuominen, J., Hyvönen, E.: Publishing and using ontologies as mash-up services. In: Proceedings of the 4th Workshop on Scripting for the Semantic Web (SFSW2008), 5th European Semantic Web Conference 2008 (ESWC 2008). (June 1-5 2008)
12. 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 ESWC 2007, Innsbruck, Austria, Springer–Verlag, Berlin (June 4–5 2007)
13. Hyvönen, E., Saarela, S., Viljanen, K.: Application of ontology techniques to view-based semantic search and browsing. In: The Semantic Web: Research and Applications. Proc. of the 1st European Semantic Web Symposium (ESWS 2004). (2004)
14. Brooke, J.: SUS: a 'quick and dirty' usability scale. In: Usability Evaluation in Industry. Taylor & Francis, London (1996) 189–194

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