HealthFinland —Finnish Health Information on the Semantic Web

Eero Hyvönen, Kim Viljanen, and Osma Suominen

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

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 publishers'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 health organizations in Finland. The system is based on a shared semantic metadata schema, ontologies, and mash-up ontology services. The content includes 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 independent 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 heterogenous environment creates an obstacle for citizens concerned about their health. Portals try to ease these problems by collecting content into a single site [1]. Portal types include *service portals* collecting a large set of services together into a localized miniature version of the web (e.g., Yahoo! and other "start pages"), *community portals* [2] acting as a virtual meeting place of a community, and *information portals* [3] acting as hubs of data. This paper discusses problems concerning information portals 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 organizations, based on a shared metadata schema, ontologies, and mash-up ontology services, by which the content is created cost-effectively by independent content producers at different locations. Our system aggregates and makes the content semantically interoperable 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

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

Information of the Semantic Web"[2]. The content for the prototype (ca. 6000 web documents) was created by the National Public Health Institute (KTL)[3], the UKK Institute [4], the Finnish Institute of Occupational Health[5], the national Suomi.fi citizen's portal[6] portal, and the Ministry of Justice[7], and new organizations are joining in.

In the following, problems of finding and producing health information on the web are first outlined. After this the content creation model of HEALTHFINLAND and the portal itself are presented.

2 Problems of Mediating Health Information

A citizen searching for health information on the web faces many challenges:

1. *Content discovery*. The discovery of relevant content is difficult because it often requires prior knowledge of the administrative organization providing the contents.
2. *Outdated and missing linkage*. After finding a piece of interesting information, it is often tedious and difficult to find related relevant web resources. Furthermore, when useful links are given on a web page, they outdate quickly. When new information is entered in a site or old information changed or removed, the links in existing pages cannot be updated automatically but refer to older information, or even non-existing information.
3. *Content aggregation*. Satisfying an end-user's information need often requires *aggregation* of content from several information providers, which is difficult if heterogeneous content is provided by several independent web sites. For example, if a baby is born in your family, relevant information related to the situation may be provided by health care organizations, social organizations, the church, legal administration, and others.
4. *Quality of content*. The trustworthiness of the information on the web pages 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.
5. *Matching end-user's expertise level.* There are lots of medical information available that is targeted to experts rather than ordinary citizens. Providing and finding the information on the right level of user expertise is a challenge that is very evident in the medical domain where, e.g., the terminology used by doctors and content providers is very different from the terminology used by citizens in expressing their needs and interests.

From the viewpoint of the health organizations, creating health information to citizens is problematic in many ways:

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

1. *Duplicated work.* Several organizations create overlapping content, which is in many cases a waste of time and money and confusing to the end-user. For example, in Finland the governmental citizen portal Suomi.fi has a section for governmental health information containing material partly overlapping with those available through the sites of the Finnish Centre of Health Promotion, and the health pages of the national broadcasting company YLE. These organizations share the goal of providing free health information to citizens and are not competing with other. In our vision, similar content should in such situations be by created only once and re-used rather than re-created by others.
2. *Difficulty of reusing content*. Content in portals is usually annotated for the purpose of presenting it in a particular portal and for the particular purpose of the organization managing the portal. 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. For example, in our case, a newspaper would be willing to publish links to the governmental HEALTHFINLAND portal to enrich their health related news articles, and the portal would definitely like to promote its health information to the readers of the online newspaper. However, a cost-effective way to do this with minimal changes in current content management systems (CMS) is needed.
3. *Internal and external link maintenance.* The problems of maintaining links up-to-date is very costly and tedious from the site maintenance viewpoint, especially when dealing with links to external sites to which the maintainer and the CMS system has no control.
4. *Indexing (annotation) problems.* 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[8], UMLS[9] or SNOMED CT[10], are very large and require expertise to use.
5. *Quality control.* 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, is the content readable and written for the correct audience).3) Quality of additional information on pages (e.g., it is advisable to show the date of publication on each page). 3) 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.

Much of the semantic web [4, 5] content will be published using semantic portals [6] based on web standards such as RDF[11] and OWL[12]. In MUSEUMFINLAND[13] [7, 8], a semantic web model and portal was created in the cultural domain for distributed semantic content creation [9], aggregation, and provision to end-users using semantic

[8] http://www.nlm.nih.gov/mesh/
[9] http://umlsinfo.nlm.nih.gov
[10] http://www.snomed.org/snomedct/
[11] http://www.w3.org/RDF/
[12] http://www.w3.org/TR/owl-features/
[13] http://www.museosuomi.fi/

search and browsing services. This approach has been shown to be applicable in different domains [1, 10], and it was also applied to HEALTHFINLAND. In the following, we show how HEALTHFINLAND develops the idea of semantic portals further and applies it in practice to create a national publication channel for health information targeted to citizens.

3 Overview of the HEALTHFINLAND 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 used to aggregate related material collections within one site, and to provide local search and linking services. Linking between sites is usually done manually. Search engines are used to provide content aggregation services on the global cross-site level.

In HEALTHFINLAND we wanted to create a new kind of collaborative distributed content creation model for publishing health information on the web in order to solve the problems listed in section 2.

The first idea of the model is 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 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 published at any moment by any of the content providers, and the system should be 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 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 providers in indexing the content.

Figure 1 depicts an overview of the HEALTHFINLAND system. The content providers on the left produce web pages, documents, and other resources of interest along their organizational interests as before for their own purposes ("primary applications" in the figure). 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 knowledge base (center of the figure) to be re-used in "secondary applications". In this paper,

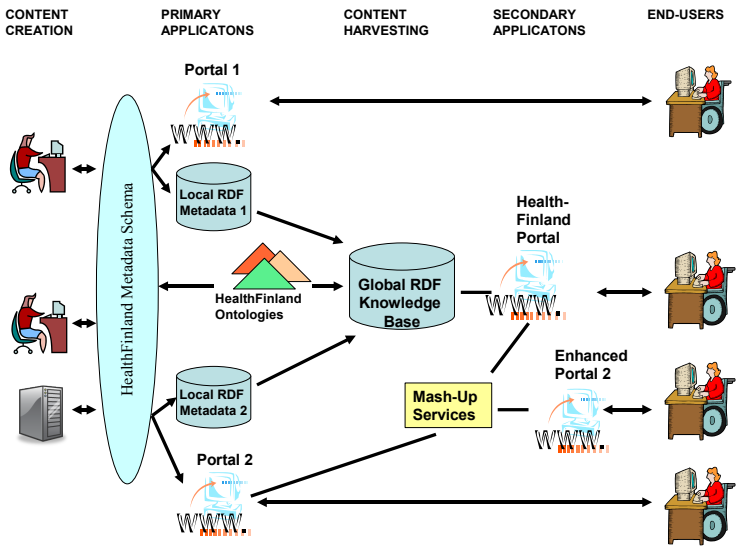

Fig. 1. An overview of the HEALTHFINLAND content creation and reuse process.

we focus on one application in particular, the semantic portal HEALTHFINLAND that provides citizens with information services to the global health information repository. We will also briefly show how external organizations can re-use the semantic content cost-effectively with semantic mash-up services called "floatlets" in the sprit of Google Maps[14] and AdSense[15], but generalized on the semantic level. The figure depicts an enhanced portal "Portal 2" in which the content of the primary application is enriched by, e.g. semantic recommendation links to content pages in HEALTHFINLAND.

In the following, the metadata schema and ontologies used in the system are first outlined.

4 Ontological Infrastructure

The ontological infrastructure of HEALTHFINLAND consists of two major components: 1) A metadata schema, i.e. an annotation ontology 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 *syntactic interoperability* of the content. 2) A set of ontological vocabularies whose concepts are used to fill in values of the metadata schema. Also the ontologies are shared by the organizations, and their usage ensures *semantic interoperability* of the content.

[14] http://maps.google.com
[15] http://www.google.com/adsense/

4.1 Metadata Schema

The HEALTHFINLAND portal requires the web documents used in the system to be described in a uniform and machine-understandable manner. A metadata schema specifies a set of fields (properties) which are used for presenting information about each document. The values of the metadata fields are either human-readable text (e.g., title), structured strings (e.g., publication date) or shared, explicitely identified ontological concepts (e.g., the subject classification). Some fields are obligatory and some fields may exists more than once. In addition to being a formal specification of what is required from the content producers, the schema can be used for, e.g., automatically generating a user interface for creating metadata conforming to the schema, and for automatic content validation and feedback generation before publishing the content in the portal [11].

The metadata schema (see table 1) is based on the Dublin Core Element Set[16], along with refinements introduced in DCMI Terms[17]. In addition, to allow a more detailed description of the required metadata, we have introduced three extensions to Dublin Core

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

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

Table 1. HEALTHFINLAND Metadata Schema. Obligatory fields are marked in **bold**. Cardinalities are presented in the column C.

	Name	QName	C	Value type	Value range
General metadata	**Identifier**	**dc:identifier**	1	URI	
	Locator	ts:url	0..1	URL	
	Title	**dc:title**	1[a]	Free text	Non-empty string.
	Abstract	**dcterms:abstract**	1[a]	Free text	Non-empty string.
	Language	**dc:language**	1..*	String	RFC 3066
	Publication time	**dcterms:issued**	1	String	W3CDTF (ISO 8601)
	Acceptance time	dcterms:dateAccepted	0..*	String	W3CDTF (ISO 8601)
	Modification time	dcterms:modified	0..*	String	W3CDTF (ISO 8601)
	Publisher	**dc:publisher**	1..*	Instance	foaf:Organization
	Creator	dc:creator	0..*	Instance	foaf:Organization, foaf:Person or foaf:Group
Content classification	**Subject**	**dc:subject**	1..*	Concept	YSO, MeSH and HPMulti Ontologies
	Audience	**dcterms:audience**	1..*	Concept	Audience Ontology
	Genre	**ts:genre**	1..*	Concept	Genre Ontology
	Presentation type	**dc:type**	1..*	Concept	DCMI Type vocabulary
	Format	**dc:format**	1	String	IANA MIME types
	Medium	**dcterms:medium**	1	Concept	Medium Ontology
	Spatial coverage	dcterms:spatial	0..*	String or concept	DCMI Point, DCMI Box or Location Ontology
	Temporal coverage	dcterms:temporal	0..*	String or concept	W3CDTF, DCMI Period or Time Ontology
Relations	Part of	dcterms:isPartOf	0..*	Document	URI
	Rights	dc:rights	0..*	Free text or document	URI or textual description
	Source	dc:source	0..*	Free text or document	URI (e.g., ISBN) or bibliographical reference
	Reference	dcterms:references	0..*	Free text or document	URI (e.g., ISBN) or bibliographical reference
	Translation of	ts:isTranslationOf	0..*	Document	URI
	Format of	dcterms:isFormatOf	0..*	Document	URI

[a] Multilingual values are allowed, but only one value in each language.

elements: 1) The *dc:type* field has been refined with a *ts:genre* field[18] to distinguish between the technical type of the document (presented using DCMI Type vocabulary) and the content genre, such as *News item*, *Organizational information* and *Research* (described in our Genre ontology). 2) The *dc:identifier* is extended with an (optional) *ts:url* field to distinguish between non-accessible identifiers and document locators. 3) The final extension is the *ts:isTranslationOf* field which extends the *dcterms:isVersionOf*, and is used for presenting the relation between language translations of documents. The metadata schema is specified in detail in [12].

Table 2. Examples of how metadata is presented in RDF/XML and XHTML.

	RDF/XML	XHTML
Free text	`<dc:title>Rokotteiden hävittäminen</dc:title>`	`<meta name="DC.title" content="Rokotteiden hävittäminen" />`
String	`<dc:language><dcterms:RFC3066> <rdf:value>fi</rdf:value> </dcterms:RFC3066></dc:language>`	`<meta name="DC.language" scheme="DCTERMS.RFC3066" content="fi" />`
Concept	`<dc:subject rdf:resource= "http://www.yso.fi/onto/yso/p123" />`	`<link rel="DC.subject" href="http://www.yso.fi/onto/yso/p123" />`

The metadata in HEALTHFINLAND is intended to be presented using RDF, conforming to the recommendations for expressing Dublin Core in RDF [13, 14]. A subset of the metadata can also be embedded in (X)HTML pages using META and LINK elements based on the Dublin Core recommendation [15]. The HTML embedded metadata solution has some limitations, because not all relevant documents are in HTML format and advanced RDF metadata structures, such as defining an instance with a certain URI, can not be done using the HTML META and LINK tags. Therefore, the RDF presentation is recommended. Examples of how metadata is expressed in RDF and HTML is shown in table 2.

The RDF and/or HTML embedded metadata is published for the HEALTHFINLAND portal by making it available on a public WWW server where it can be accessed regurlarly by the HEALTHFINLAND metadata harvester which fetches the content from the content providers to a centralized metadata server (cf. Figure 1). 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 that were 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. Otherwise, the metadata is added to and published in the HEALTHFINLAND portal.

[18] namespace *ts* refers to the Finnish name TerveSuomi of HEALTHFINLAND

4.2 Ontologies

Semantic interorerability in HEALTHFINLAND 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 (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 genre types (news, game, 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)[19] that includes approximately 20 000 concepts. The YSO ontology was created by transforming the General Finnish Thesaurus YSA[20] into RDF/OWL format using the Protégé editor[21] and by manually crafting the concepts into full-blown rdfs:subClassOf hierarchies [16]. YSA is widely used in Finland for indexing various kinds of content, e.g. in libraries.
2. The international Medical Subject Headings (MeSH) which includes approximately 23 000 concepts. The Finnish translation of MeSH, FinMeSH, was developed by the Finnish Medical Society Duodecim[22] and was acquired for HEALTHFINLAND as a database. The vocabulary was transformed into the SKOS Core format[23] without changing the semantics of the vocabulary or its structure.
3. The European Multilingual Thesaurus on Health Promotion [24] (HPMULTI), which included a Finnish translation. HPMULTI contains approximately 1200 concepts related specifically to health promotion. HPMULTI was transformed into SKOS/RDF in the same way as FinMeSH.

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 clinical healthcare. HPMULTI complements the two vocabularies by focusing on health promotion terminology.

5 Distributed Semantic Content Creation

A major challenge in the distributed content creation model of HEALTHFINLAND is how to facilitate the cost-effective production of descriptive, semantically correct high-quality metadata. In HEALTHFINLAND three ways of creating metadata are considered

[19] http://www.seco.tkk.fi/ontologies/yso/
[20] http://www.vesa.lib.helsinki.fi
[21] http://protege.stanford.edu
[22] http://www.duodecim.fi
[23] http://www.w3.org/2004/02/skos/core/
[24] http://www.hpmulti.net/

and supported: 1) Boosting existing web content management systems (CMS) with ontology mash-up services for producing semantic metadata. 2) Using a browser-based metadata editor for annotating web content. 3) Automatical conversion of metadata. These approaches are explained shortly below.

5.1 Boosting an Existing CMS with Mash-Up Ontology Services

Most content providers in HEALTHFINLAND use a CMS for authoring, publishing and archiving content on their website. A typical CMS systems supports creation of textual metadata about documents, such as title and publication time, but not ontological annotations. This would require that the system has functionalities supporting ontology-based annotation work, e.g., concept search for finding the relevant concepts (identified with URIs), concept visualisation for showing the concept to the user, and concept storing along other information about the documents. The CMS should also be able to export the metadata preferably in RDF format to be used by semantic web applications.

Currently, ontologies are typically shared by downloading them, and each application must separately implement the ontology support. To avoid duplicated work and costs, and to ensure that the ontologies are always up-to-date, we argue that one should not only share the ontologies, but also the *funtionalities* for using them as centralized mash-up services. Such services, e.g. Google Maps, have been found very useful and cost-effective in Web 2.0 applications for integrating new functionalities with existing systems.

We have applied the idea of using mash-ups to provide ontology services for the content producers of HEALTHFINLAND by creating the ONKI Ontology Server framework[25] [17]. ONKI provides ontological functionalities, such as concept searching, browsing, disambiguation, and fetching, as ready-to-use mash-up components that communicate asynchronously by AJAX[26] (or Web Service technologies) with the shared ontology server. The service integration can be done easily by changing only the user-interface component slightly at the client side. For example, in the case of AJAX and HTML-pages, only a short snippet of Java Script code must be added to the web page for exploiting the ONKI services.

The main functionality addressed by the ONKI UI components is concept finding and fetching. For finding a desired annotation concept, ONKI provides text search with semantic autocompletion [18]. This means that when the annotator is typing in a string, say in an HTML input field of a CMS system, the system dynamically responds after each input character by showing the matching concepts on the ONKI-server. By selecting a concept from the result list, the concept's URI, label or other information is fetched to the client application.

Also concept browsing can be used for concept fetching. In this model, the user pushes a button on the client application that opens a separate ONKI Browser window in which annotation concepts and be searched for and browsed. For each concept entry, the browser shows a *Fetch concept* button which, when pressed, transfers the current concept infromation to the client application.

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

[26] http://dojotoolkit.org/

ONKI also supports multilingual ontologies, has a multilingual user-interface, supports loading multiple ontologies, and can be configured extensively.

ONKI is implemented as a Java Servlet application running on Apache Tomcat. It uses the Jena semantic web framework for handling RDF content, the Direct Web Remoting (DWR) library for implementing the AJAX functionalities, the Dojo Javascript toolkit, and the Lucene text search engine.

5.2 Browser-based Metadata Editor

Some HEALTHFINLAND content providers can not add mash-up ontology support to their CMS due to technical or economical reasons. Furthermore, some content providers do not even have a CMS or they may not have access to the CMS that contains the content, e.g., if the content originates from a third party. To support metadata productions in these cases, we have created a centralized browser-based annotation editor SAHA [11] for annotating web pages. SAHA adapts automatically to different metadata schemas. In this case the HEALTHFINLAND schema is used. The schema element fields in SAHA can be connected with ONKI mash-up ontology services, providing concept finding and fetching services to the annotator, as discussed above.

5.3 Automatical Conversion

The third content producing method in HEALTHFINLAND is automatical conversion of original data to HEALTHFINLAND metadata. This method is used currently in cases where metadata exists in a CMS, but it is in an incompatible format, does not contain ontological annotations (URIs) and/or some minor information is missing in the metadata. Because the HEALTHFINLAND metadata schema is strongly based on Dublin Core and because many content providers in Finland use thesauri (e.g., the Finnish General Thesaurus YSA and the Medical Subject Headings MeSH), the content in many cases can be transformed fairly accurately into ontological form automatically. For example, some legal content produced by the Finnish Ministry of Justice is harvested for HEALTHFINLAND. The metadata, targeted originally for the govermental Suomi.fi portal[27], uses a Dublin Core based metadata schema (JHS 143 recommendation [19]) and is automatically translated into the HEALTHFINLAND metadata format.

6 Intelligent Services to the End-User

The HEALTHFINLAND user interface is based on the faceted browsing (a.k.a. view-based search) paradigm [20, 21], which has been found useful in our earlier semantic portals, such as [7, 1, 10], and in other systems, such as SWED[28] and MultimediaN[29].

A challenge in publishing health-related information in a citizens' semantic portal is the gap between the citizens' information needs and the professional conceptualizations

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

[28] http://www.swed.org.uk/swed/index.html

[29] http://e-culture.multimedian.nl/demo/search

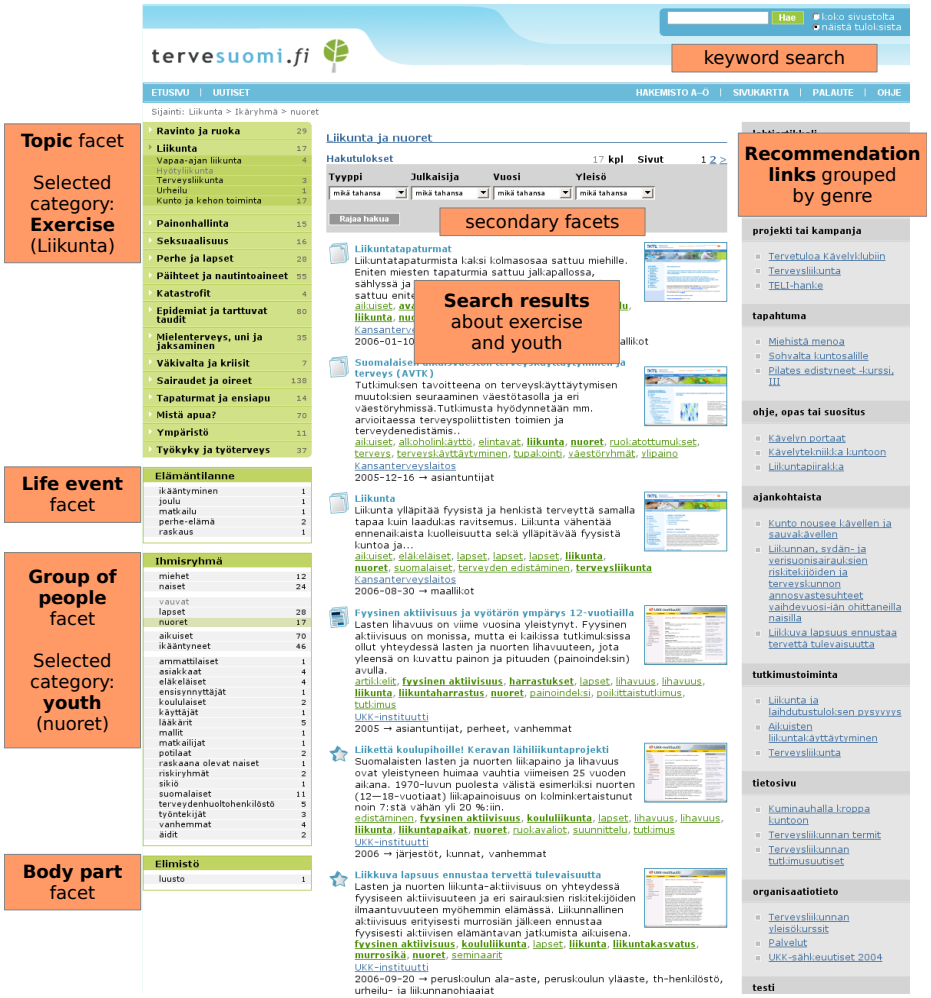

Figure 2. Portal user interface with semantic search, browsing and recommendations

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 [22] to elicitate 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 HEALTHFINLAND portal, like typical semantic portals, provides the end-user with two basic services: 1) a search engine based on the semantics of the content and 2) dynamic linking between pages based on the semantic relations in the underlying knowledge base. The main facets of the portal are Topic, Life event, Group of people,

and Body part. The facets can be seen in the left column in figure 2. 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.

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.

The portal also provides 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).

One problem with a portal approach and distributed content creation in general is that when search results are provided as traditional hyperlinks, users are forced to navigate between different web sites that each have their own navigation systems and styling. Also, providing recommendation links across sites is challenging.

The HEALTHFINLAND portal will integrate selected content items that have been retrieved from affiliated websites directly into the portal interface, providing seamless navigation and recommendation links in the proper context of the content page. Our solution requires that the content is marked up using a small amount of RDFa syntax[30], which helps the metadata harvester extract the body content of suitable web pages, skipping navigation elements and styling.

The HEALTHFINLAND 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. 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.

7 Discussion

This paper addressed the problems of the citizen end-users (cf. section 2) 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 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 indexing vocabularies makes it possible to the

[30] http://www.w3.org/TR/xhtml-rdfa-primer/

citizen search and browse content using in layman's vocabulary although the content is indexed by professional medical terminology .

At the same time, the problems of content providers (cf. Section 2) are eased, too: 1) Duplication of content creation can be minimized by the possibility of aggregating cross-portal content. 2) Reusing the global content repository is feasible, as demonstrated by the semantic portal HEALTHFINLAND. By using mash-up floatlets, external applications, such as the primary applications of figure 1, can reuse the content provided by secondary applications, such as HEALTHFINLAND. 3) Internal and external link managament 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 ontology-values metadata schema as in [23]. However, the idea of sharing ontologies through mash-up ontology services in a distributed environment is new. The user interface is based on the faceted search paradigm [20, 21], but integrated with semantic web ontologies and reasoning with semantic recommendations [24], as in [25]. A new feature of the system is the separation of end-user facets from indexing ontologies [26, 22], which is crucial in the medical domain. The card sorting approach [22] was found useful in accomplishing this.

This work is a part of the national semantic web ontology project FinnONTO[31] 2003-2007, 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. 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. Staab, S., others, J.: Semantic Community Web Portals. In: Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, Elsevier (2000)
3. Reynolds, D., Shabajee, P., Cayzer, S.: Semantic Information Portals. In: Proceedings of the 13th International World Wide Web Conference on Alternate track papers & posters, New York, NY, USA, ACM Press (2004)
4. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **284**(5) (2001) 34–43
5. Fensel, D.: Ontologies: Silver bullet for knowledge management and electronic commerce (2nd Edition). Springer-Verlag (2004)
6. Maedche, A., Staab, S., Stojanovic, N., Struder, R., Sure, Y.: Semantic portal — the SEAL approach. Technical report, Institute AIFB, University of Karlsruhe, Germany (2001)

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

7. 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
8. 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: Semantic Web at Work - Proceedings of the 12th Finnish Artificial Intelligence Conference STeP 2006, Volume 1. (2006)
9. Hyvönen, E., Saarela, S., Viljanen, K., Mäkela, E., Valo, A., Salminen, M., Kettula, S., Junnila, M.: A semantic portal for publishing museum collections on the web. In: Proceedings of ECAI/PAIS 2004, Valencia, Spain. (2004)
10. Känsälä, T., Hyvönen, E.: A semantic view-based portal utilizing Learning Object Metadata (2006) 1st Asian Semantic Web Conference (ASWC2006), Semantic Web Applications and Tools Workshop.
11. 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, Berlin (2007)
12. 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/.
13. Dublin Core Workgroup: Expressing simple Dublin Core in RDF/XML (2002) http://dublincore.org/documents/dcmes-xml/.
14. Dublin Core Workgroup: Expressing qualified Dublin Core in RDF/XML (2002) http://dublincore.org/documents/dcq-rdf-xml/.
15. Dublin Core Workgroup: Expressing Dublin Core in HTML/XHTML meta and link elements (2003) http://dublincore.org/documents/dcq-html/.
16. 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)
17. Hyvönen, E., Viljanen, K., Mäkelä, E., et al.: Elements of a national semantic content infrastucture—case finland on the semantic web. (2007) submitted paper under evaluation.
18. Hyvönen, E., Mäkelä, E.: Semantic autocompletion. In: Proceedings of the first Asia Semantic Web Conference (ASWC 2006), Beijing, Springer-Verlag, New York (2006)
19. JHS workgroup: JHS 143: Asiakirjojen kuvailun ja hallinnan metatiedot (2004) http://www.jhs-suositukset.fi/suomi/jhs143.
20. 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.
21. 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
22. 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)
23. Hyvönen, E., Salminen, M., Kettula, S., Junnila, M.: A content creation process for the Semantic Web (2004) Proceeding of OntoLex 2004: Ontologies and Lexical Resources in Distributed Environments, May 29, Lisbon, Portugal.
24. 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)
25. 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)

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