HealthFinland —a National Semantic Publishing Network and Portal for Health Information

Osma Suominen [a,*], Eero Hyvönen [a], Kim Viljanen [a], Eija Hukka [b]

[a] *Semantic Computing Research Group (SeCo),*
Helsinki University of Technology (TKK), Department of Media Technology
University of Helsinki, Department of Computer Science
http://www.seco.tkk.fi
[b] *National Institute for Health and Welfare (THL)*
http://www.thl.fi

Abstract

Providing citizens with reliable, up-to-date and individually relevant health information on the web is done by governmental, non-governmental, business and other organizations. Currently the information is published with little co-ordination and co-operation between the publishers. For publishers, this means duplicated work and costs due to publishing same information twice on many websites. Also maintaining links between websites requires work. From the citizens point of view, finding content is difficult due to e.g. differences in layman's vocabularies compared to medical terminology and difficulties in aggregating information from several sites. To solve these problems, we present a national scale semantic publishing system HealthFinland which consists of a 1) a centralized content infrastructure of health ontologies and services with tools, 2) a distributed semantic content creation channel based on several health organizations, and 3) an intelligent semantic portal aggregating and presenting the contents from intuitive and health promoting end-user perspectives for human users as well as for other websites and portals.

Key words: health information, semantic portal, metadata, faceted search, content aggregation, ontology service

1. Introduction

Health information is one of the most frequently searched materials on the web. Among American Internet users, about 8 million adults searched for information on at least one health topic on a typical day in August 2006, most commonly information about specific diseases and problems, treatments, diet and nutrition or exercise [26]. However, a citizen searching for health information on the web faces many challenges [17].

General web search engines are the most common starting points used to find health information [26]. Web search engines require that users know or can guess suitable keywords. Search engines are generally not aware of the context of use and cannot match content to user needs based on, e.g., target audience or other demographic information. Apart from information that search engines can infer from link structure and other statistical measures, there

* Corresponding author. Tel: +358 9 451 3346
Email addresses: osma.suominen@tkk.fi (Osma Suominen), eero.hyvonen@tkk.fi (Eero Hyvönen), kim.viljanen@tkk.fi (Kim Viljanen), eija.hukka@thl.fi (Eija Hukka).

is no notion of trusted information sources or quality measures relevant to assessing the trustworthiness of health information and advice.

In practice, general web search result listings for common health topics often contain a heterogeneous mix of health-related webpages. As an example, the top twenty search results for `diabetes` retrieved by Google on April 21, 2009 include news sites, non-governmental organizations such as patient groups, a Wikipedia article, commercial sites that promote health-related products and services as well as government sites that seek to inform citizens. More specific queries such as `diabetes pregnant` also turn up user-generated content, e.g., discussion forum postings, blogs and video clips.

Sorting out relevant, reliable and useful content from this mix, and assessing the quality of health websites, requires background knowledge about the organizations that publish health information on the web. When users do find interesting resources, finding related relevant content is not very easy as no single website can provide everything a user may be interested in, and the linking between sites is often insufficient. Many common information needs and situations, such as when a baby is born into a family, require *aggregation* of content from several information providers.

The quality and trustworthiness of information varies. In many cases, it is difficult to know whether information on a web page is based on scientific research, layman opinions and rumors, or whether it is motivated by commercial interests. Even basic quality indicators such as information sources and publication dates are often absent on health websites, and users do not consistently check for them [26]. Even when good quality information is found, it may not be targeted for the right audience: a treatment recommendation for gestational diabetes issued by a government agency for doctors may be useless for an anxious future mother.

From the health information publishers' point of view, a central problem today is that there are many actors and websites but little coordination and cooperation between them, resulting in duplicate work even between non-competitive organizations such as different government institutions [17]. Reusing pre-existing web content is mainly done by manually linking websites, but such links quickly expire when websites are restructured or shut down. Creating good quality health content is problematic and requires a process for quality control, e.g., regular reviews and updates.

The annotation of web content with metadata is not very common on health websites, but it is sometimes mandated by, e.g., government standards. In these cases, the annotation processes are often tedious and require expertise in the use of large vocabularies such as the Medical Subject Headings (MeSH) [1].

HealthFinland [2] is a semantic health information publishing system that addresses these problems both from the publishers' and citizens' viewpoints. The concept and ideas behind HealthFinland have been presented in [17,18]. The design and implementation of the portal information architecture and user interface was presented [33], and the design and implementation of the portal in more detail in [32]. The FinnONTO ontology service infrastructure and services used in the system are presented in [42,19,36]. This article complements our earlier work and publications by providing a more detailed description of the HealthFinland portal components, the content creation infrastructure, the user interface, and the evaluations performed. Based on the lessons learned in developing the prototype, the system has been developed further and is being published in a production environment as a national health portal.

In the following, we first present and overview of the ideas underlying HealthFinland and the major components and architecture of the system. We then present the ontologies and other domain vocabularies used in the system as well as the metadata model and the underlying content production system. Finally, we present the user interface of the semantic portal. In each section, lessons learned in designing and implementing the prototype are discussed, and related work is pointed out. In conclusion, contributions of the research are summarized and directions for further research are outlined.

2. The HealthFinland Approach

The main reasons for developing HealthFinland using Semantic Web technologies are [17]: facilitating cost-effective, distributed content creation in an interoperable way, aggregating contents automatically based on semantics, and providing the end-users with intelligent services.

First, HealthFinland minimizes duplicate redundant work and costs in creating health content

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

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

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 has the best professional experience about the topic. 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 website. 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.

The second key idea behind HealthFinland is to try to minimize the content 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 is 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 HealthFinland is to provide the end-user with intelligent services for finding the right information based on his or her own information needs or 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 a "layman's" vocabulary that is different from the medical expert vocabularies used by the content providers in indexing the content [33].

To implement these ideas, HealthFinland has three main components (cf. Figure 1): 1) The Domain Vocabularies describe the different aspects of health information such as topics, genres and audiences, and they are published in an ontology repository as part of the FinnONTO system of mutually interlinked vocabularies/ontologies [19]. 2) The Content Production System contains specifications and tools for annotating, harvesting, and verifying content. 3) The Semantic Portal for providing the contents for human users through semantic search and browsing services and for machines, i.e., other web applications, as semantic widgets (called "floatlets") [23].

Figure 2 gives a more detailed overview of the HealthFinland system architecture. Metadata and documents are collected from the content publishers either by harvesting content and metadata from their content management systems or by annotating content manually using the SAHA metadata editor [37] connected to ONKI ontology services [36]. The content is validated and possible problems are reported to the content providers. The successfully validated metadata is finally published through the portal for humans and machines to use.

3. Domain Vocabularies

A semantic portal needs a set of vocabularies which are referenced by the metadata. The ontologies should describe all the relevant concepts in the application domain of the portal. In HealthFinland ontologies and other controlled vocabularies are used for describing document subject matter, content genres and target audiences.

3.1. *Subject Vocabularies*

In the case of health information and health promotion, the subject matter ontologies must describe a large number of possible topics such as diseases, treatments and anatomy as well as living habits such as diet, exercise and substance use.

We started by analyzing pre-existing, established ontologies and other structured vocabularies for describing health-related content. Reusing existing ontologies offers several advantages: it eases semantic interoperability with other applications, saves time and money by avoiding unnecessary ontology engineering work, and helps to ensure broad coverage of the subject area as established ontologies typically have been used and developed for a long time. We settled on three *core subject domain* ontologies that are used for describing the subject matter of web contents:

(i) The Finnish General Upper Ontology (YSO) [3] that includes approximately 20 000 concepts. The YSO ontology was created by transforming the General Finnish Thesaurus YSA [4] into RDF/OWL format using the Protégé editor [5] and by manually crafting the concepts into *rdfs:subClassOf* hierarchies [19].

(ii) The international Medical Subject Headings (MeSH) which includes approximately 23 000

[3] http://www.seco.tkk.fi/ontologies/yso/
[4] http://www.vesa.lib.helsinki.fi
[5] http://protege.stanford.edu

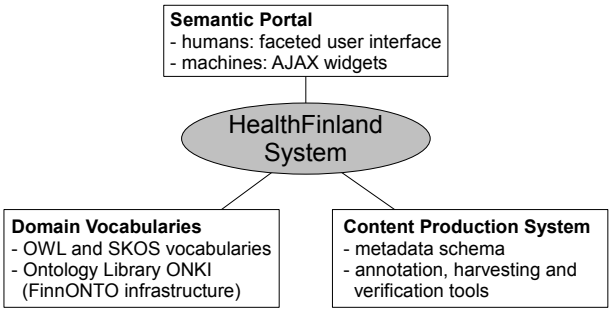

Fig. 1. Three main components of HealthFinland.

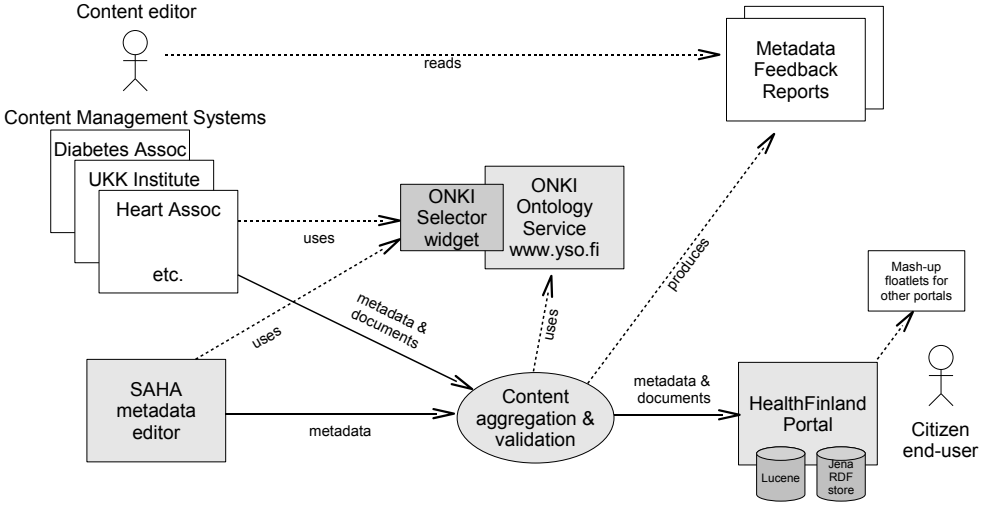

Fig. 2. HealthFinland system architecture.

concepts. We transformed the vocabulary into the SKOS Core format [6] without changing the semantics of the vocabulary or its structure using conversion tools and methods developed at the Free University of Amsterdam [38].

(iii) The European Multilingual Thesaurus on Health Promotion [7] (HPMULTI), which included a Finnish translation. HPMULTI contains approximately 1200 concepts related specifically to health promotion. We transformed HPMULTI into SKOS/RDF in a similar way as MeSH using a custom conversion tool.

All three vocabularies are required to cover the subject matter of the portal properly. YSO is broad but does not include enough concepts that describe 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 specific terminology.

To make the three subject vocabularies semantically compatible with each other, we created the Health Promotion Ontology using YSO as the structural basis and extended it manually, using the Protégé ontology editor, 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.

3.2. *Content Genre Vocabulary*

Content genres describe different types of documents based on purpose, form and content [29]. Genres can be specific to the application domain, and in the case of health information, some impor-

[6] http://www.w3.org/2004/02/skos/
[7] http://www.hpmulti.net/

tant genres include *articles*, *tests* (for assessing your current health or habits), *official recommendations* and *news*. Although genres are currently not very widely used in web search, they provide an interesting means of restricting content searches in addition to subject- or keyword-based searching.

We could not find a suitable pre-existing genre vocabulary for health information, so we used a lightweight user-centered process very similar to Rosso's card sorting survey method [29] to construct a classification of genres. Five test subjects (Rosso's experiment had only three participants in the initial phase) were asked to sort a selection of forty printed content pages into groups based on their form and structure, and finally label them with descriptive names. An initial genre vocabulary was constructed based on an analysis of the results. The vocabulary was later extended and restructured as new genres were identified that were not well represented in the original experiment. Some categories were merged or renamed based on experiences from user testing. The resulting genre classification currently includes 21 content genres and is expressed as a SKOS Core vocabulary.

3.3. *Target Audience Vocabulary*

Describing the target audience of documents allows content to be selected and promoted to certain users. Audience-based matching requires a vocabulary of audiences and a user interface that makes use of audience information.

To describe the different target audiences of the HealthFinland portal, we initially created a Target Audience Ontology. We constructed the ontology by taking into account five possible ways of segmenting audiences: by age, gender, occupation, and patient group as well as some other special groups and roles, e.g., parents and exercise teams. We used the classification of occupations published by Statistics Finland [8] and a smaller target audience classification used in a publication database at the UKK Institute [9] as starting points for the ontology. We organized the different occupations based on a division into health professionals (e.g. doctors, nurses and therapists) and other, non-health occupations. The ontology was built using the Protégé ontology editor and transformed into SKOS Core format. The final ontology contains more than 800 target audiences.

However, we soon discovered that the size and structure of the ontology was unwieldy for annotations and the hierarchies often unsuitable for processing search queries in a health portal. The occupation classification we used was intended for a different purpose (collecting national level statistics about, e.g., salaries) and mainly distinguished different occupations by the level of education they require, which had little bearing on the needs of portal audiences. Additionally, as the needs of professionals are very different from those of ordinary citizens, we decided to drop health professionals as a target audience altogether in order to better serve ordinary citizens in situations and tasks involving their personal health. Many other websites and portals that fulfil work-related needs of health professionals already exist, and new ones, such as THL.fi [10] published by the National Institute for Health and Welfare, were being developed in parallel to HealthFinland.

Finally, we settled on a relatively small audience classification that takes into account the needs and goals of users as well as the tone of discussion in the available content. The classification consists of two levels: special categories for specific situations – e.g. families with a baby, occupational health, and information for travellers – as well as three general categories for disease- and problem-oriented information, health improvement, and environmental issues, respectively. The intent is that content that fits one of the special categories is classified therein, while the rest is classified into one of the general categories. The current classification includes nine categories, but more special categories are expected to be defined when new content is added into the HealthFinland portal. The classification is expressed using the SKOS Core vocabulary.

3.4. *National Ontology Library Service ONKI*

Methods for publishing controlled vocabularies and using them in applications are required when used in professional content creation work such as in the case of the HealthFinland content creation network. At the start of the HealthFinland project, none of the Content Management Systems (CMS) used by the content publishers supported

[8] http://www.stat.fi
[9] http://www.ukkinstituutti.fi
[10] http://www.thl.fi

referencing concepts described in controlled vocabularies by URIs in the metadata, as would be required by true Semantic Web applications. Vocabularies also evolve constantly, so methods for publishing the latest versions are needed in such a way that all parties are always using the latest version of the vocabularies.

The HealthFinland content creation network uses The Finnish Ontology Library Service ONKI [11] for publishing all the domain vocabularies discussed above. ONKI is a pilot system for deploying ontologies and using them on a national scale [41]. ONKI focuses especially on ontology publishing and using them in content indexing and information retrieval through both end-user and application interfaces. ONKI ensures that all content creators use the latest versions of each ontology. From the point of view of the content creators, ONKI is used for browsing the ontologies and to support annotation of content with ontological concepts. ONKI is also used by the HealthFinland content aggregator application for finding concept URIs for the term labels.

4. Content Production System

A semantic portal requires rich metadata with well-defined object types, properties (fields) and value ranges in order to provide rich search facilities to end users. To ensure interoperability and ease integration of systems, metadata schemes should be based on common standards and frameworks such as Dublin Core [4]. However, general purpose metadata frameworks may be adapted for particular domains and applications by specifying additional constraints and extensions.

A trade-off exists between the level of semantic richness and ease of metadata creation. If annotating objects requires large amounts of manual work and the results and benefits of metadata production are not clearly visible, less metadata will likely be produced, or shortcuts will be taken that compromise metadata quality [8]. Metadata schemas should focus on effective annotations with an appropriate breadth and expressivity that does not hinder annotation while still giving benefits for the users who rely on the metadata.

We designed the HealthFinland metadata schema using the Dublin Core Metadata Element Set and the Finnish JHS143 recommendation on document metadata [20] as starting points, choosing only fields that were relevant in the domain of health information. The schema was extended with a domain-specific *Genre* field that describes health information content types such as *article*, *guide*, *news item* and *campaign*. The original schema has been published [17], and a detailed metadata specification for system implementors is available [34].

The metadata in HealthFinland is represented using RDF, conforming to the recommendations for expressing Dublin Core in RDF [6]. A subset of the metadata can also be embedded in (X)HTML pages using META and LINK elements based on the Dublin Core recommendation [5].

4.1. *Metadata Production Tools*

Creating good quality metadata requires suitable tools that assist content creators and annotators in the metadata production process. Many content producers use a content management system (CMS) to produce and publish web documents. Adding support for metadata production using the HealthFinland metadata schema into such systems makes it easy for content authors to produce metadata and keep it synchronized with updates in the content. In such a setting many metadata fields such as publication and modification dates can also be automatically assigned values.

To make it easier to assign subjects and other concept values from the vocabularies used in the HealthFinland metadata schema, we have developed the ONKI Selector Widget (Figure 3) [40]. It is an AJAX-enabled component that can easily be integrated into any HTML form and allows the selection of concepts from any vocabulary available on the ONKI server. The concepts picked using the component are stored in an HTML form field like any other metadata. This way, the content management system does not need to be aware of the vocabularies and the newest vocabulary versions are always directly available from the ONKI server.

To publish the metadata for use by the HealthFinland system, the content management system must be configured to expose its metadata either embedded into individual HTML pages or as RDF data retrievable using HTTP.

Not all content publishers use a CMS, and even if they do, adding metadata support to web publishing systems may not always be technically or economically feasible. To create metadata for such content

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

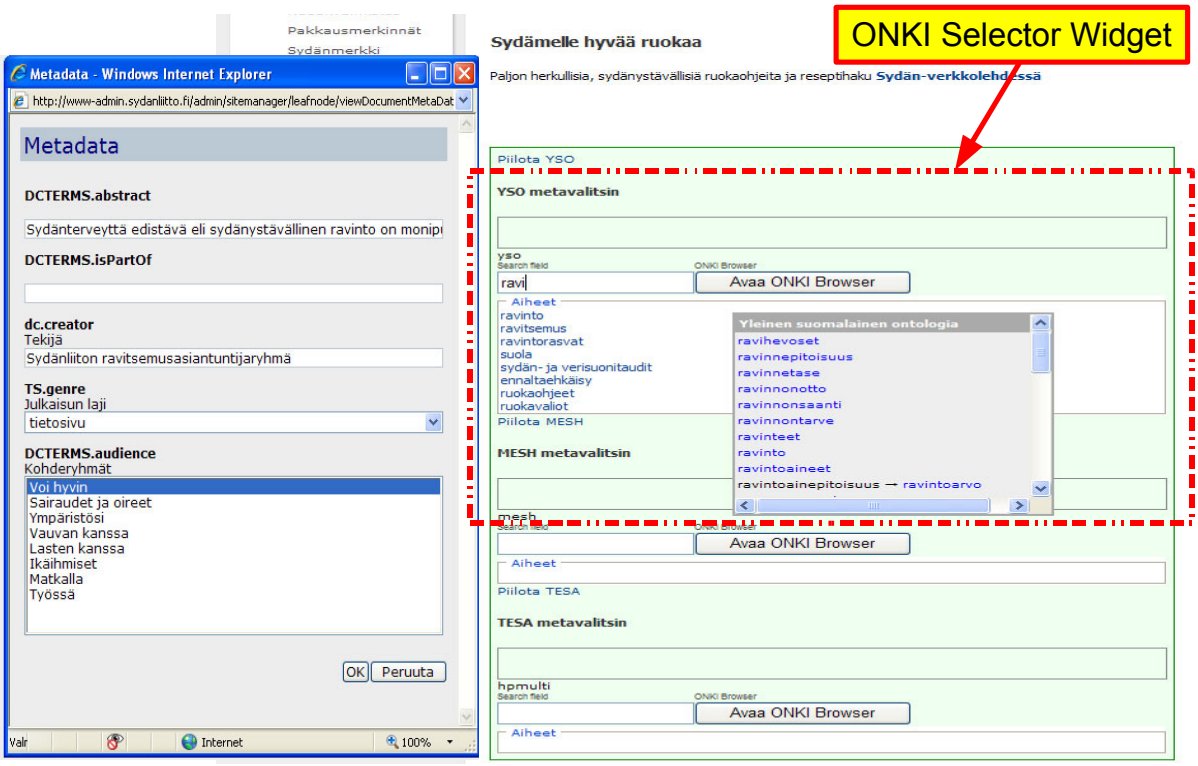

Fig. 3. Finnish Heart Association's metadata production tool using ONKI Selector Widget. The widget has been integrated into the Navigo portal software used by the association.

sources, we have developed the browser-based metadata editor SAHA [37] that can be configured with the HealthFinland metadata schema. It was used in HealthFinland for manual annotation of health information. SAHA uses the above mentioned ONKI Selector Widget for supporting concept lookup from vocabularies available on the ONKI server.

4.2. *Content Publishers*

Eight Finnish health organizations—the Finnish Diabetes Association, the Finnish Medical Society Duodecim, 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 for Health Promotion Research—contributed to the health contents available in the portal protype. The system contains almost 3500 health content objects.

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 transformers; 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 Health Promotion Ontology before the transformation could be completed.

4.3. *Metadata Aggregation and Quality Control*

Using heterogeneous data sources presents some challenges with regard to data validity and quality control. We have developed an aggregation and validation tool that collects metadata from different sources and processes it into a well-defined format for use in the HealthFinland portal. The meta-

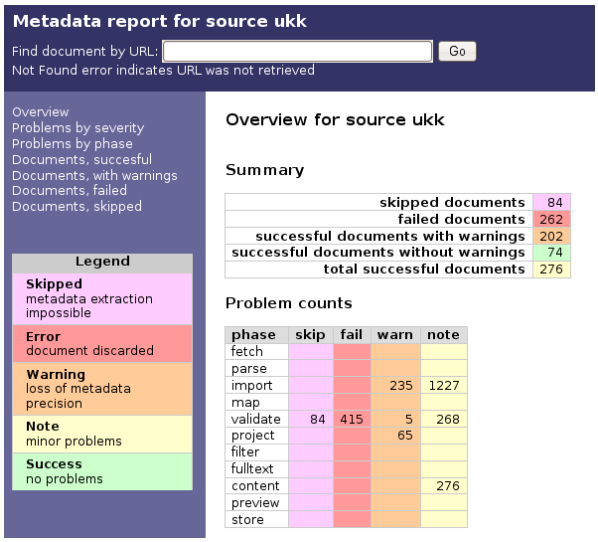

Fig. 4. Metadata Feedback Report

data is collected either by crawling the web sites of content publishers or by retrieving metadata published in RDF format. Metadata expressed as part of HTML documents is syntactically transformed into RDF, and any concepts expressed using concept labels are converted into URI references using Web Service facilities for concept lookups provided by the ONKI server. Finally, the aggregator validates the metadata against the HealthFinland metadata schema, and produces output in an RDF format that is guaranteed to conform strictly to the HealthFinland metadata schema. Any problems encountered during the aggregation and validation process are reported back to the publisher by creating a problem report, shown in Figure 4, which is implemented as a set of browsable static HTML pages.

4.4. *Experiences and Feedback*

To gain some understanding into how the initial content creation infrastructure works from the content publishers' point of view, we collected user feedback from annotators after several months of use. The feedback was collected using an e-mail survey and a meeting between the content editors where the experiences of different content producers was discussed and summarized into a list of notable problems and possible improvements. Overall, the processes and tools were considered functional, but many improvements were also suggested.

The annotators suggested that the metadata schema and related annotation tools could be improved by removing or hiding some optional fields (e.g., Coverage) that were seldom used. They also requested tools that could help them avoid repetitive work, e.g., by using previous annotations as templates for annotating new content.

The use of three different complementary subject vocabularies was considered somewhat impractical as it forced annotators to look up concepts in each vocabulary. The annotators had also encountered situations where a concept they needed was not present in any of the available vocabularies, and requested some means to add new necessary concepts into the vocabularies during the annotation work.

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

The content publishers also requested a way to suppress content from appearing in the HealthFinland portal even though valid metadata had been created. This was necessary for situations where content had been produced that was useful for the intended audience of the originating website but not relevant or useful to the general public.

4.5. *Refined Content Creation Process and Metadata Schema*

Based on the feedback, we made several improvements to the content creation process. The usability of SAHA and the problem reports has been enhanced in many ways, such as fixing user interface glitches and removing unimportant problems from the reports. The metadata schema has been updated, removing many optional fields that were not widely used in practice. The new schema allows the direct use of the Health Promotion Ontology as a subject vocabulary. The content publishers can thus avoid using all the other subject vocabularies except when they consider them necessary due to legacy metadata or requirements of direct semantic interoperability towards third party systems. However, as the Health Promotion Ontology contains mappings to the original source vocabularies, expressed using the SKOS mapping properties, indirect semantic interoperability with systems using the original vocabularies is possible regardless of which sub-

ject vocabularies the content publishers decide to use.

The new metadata schema, shown in Table 1, has been improved by removing fields that were not considered useful. Some fields, such as *Spatial* and *Temporal coverage*, *Part of* relations, *References* and *Translations*, had not been used in practice by annotators. The *Media type* field was intended to be used for distinguishing between web content and other publications such as printed books and CDs, but it was dropped as the scope of the portal search functions was restricted to web content. The *Source* field was sometimes used in the metadata, but its usage was not consistent. The inconsistencies caused some technical and presentation issues in the portal user interface, so the field was dropped from the metadata schema and publishers were instead advised to present sources in the content itself.

The definition of some metadata fields were also made stricter. In the original schema, documents could be categorized into multiple *Genres*, *Audiences*, *Presentation types* and *Languages*. After analyzing the situations where multiple categorizations for these fields had been used, we found that in most cases the content being described was not suitable for inclusion in the portal. For example, when the metadata for a document claimed that it was intended for multiple audiences, in reality it was typically not written for anyone in particular and thus would not be very helpful for any practical task faced by the users of the portal. Thus, we decided to allow a single document to be placed in only one category for each of these fields. Problematic documents that cannot obviously be classified into a single category will either be classified into the most relevant category, reworked, or dropped from the portal entirely. As a side benefit, this restriction also simplified the implementation of the portal user interface and the forms used for annotation.

Some metadata fields were also added to the schema at the request of annotators. A literal field for *Keywords* was added with the intent that annotators can use it to describe subjects that cannot be found in any of the available subject vocabularies. Frequently occurring keywords can then be considered candidates for inclusion as concepts in the Health Promotion Ontology. Additionally, a field for preventing the inclusion of content into the HealthFinland portal (*NoIndex*) was added, to give content publishers a degree of control over the publishing of their content through the HealthFinland portal.

The specification for representing metadata on web pages has also been updated to reflect the updated recommendations of the Dublin Core workgroup [5]. The representation of metadata on HTML pages now includes a reference to the Dublin Core metadata profile that, in principle, allows the metadata to be parsed and transformed into a valid RDF/XML representation by any software agent that supports the W3C GRDDL recommendation [43]. However, our content aggregator still uses a custom HTML metadata parser as it is more robust than typical GRDDL parsers in the face of syntactical problems in the HTML code and can also process legacy representations of HealthFinland metadata based on our older specifications. Our parser can also make domain-specific assumptions and corrections that a general-purpose GRDDL agent cannot be expected to make.

5. Semantic Portal

In order to fulfil its purpose, a citizens' health portal must be usable from the perspective of ordinary citizens. The usability of an information portal is affected by, e.g., the layout and functionalities of the user interface. In a semantic portal used for information retrieval, additional factors that affect usability include the search facilities available and the user-visible parts of ontologies and other vocabularies used in the portal.

The faceted search paradigm [16] was chosen as the basis for the search facilities in the HealthFinland portal, as it had been successfully applied in other semantic search systems [9,15]. In faceted search, users navigate and make search queries in the portal by making category selections along the available facets. However, finding a useful and intuitive set of facets and categorizations in the domain of health information was a challenge.

In many faceted search systems, the facets are directly constructed using the values found in metadata fields and the vocabularies that define the sets of possible field values [9,27,30]. In some systems, such as MuseumFinland, logical rules are used to define additional facets that do not directly correspond to the structure of the ontology [15].

In HealthFinland the main subject matter vocabularies have been created for and by professionals such as doctors, nurses and librarians for use at work or for scientific purposes. Their structure and choice of terms are not always appropriate for the

Table 1
HealthFinland Metadata Schema. Required fields are marked in **bold**.

Name	QName	Card.	Value type	Value range
Identifier	**dc:identifier**	1	URI	
Locator	ts:url	0..1	URL	
Title	**dc:title**	1	Literal	Non-empty string
Description	**dc:description**	1	Literal	Non-empty string
Language	**dc:language**	1	Typed literal	RFC 4646
Publication time	**dct:issued**	1	Typed literal	W3CDTF (ISO 8601)
Acceptance time	dct:dateAccepted	0..1	Typed literal	W3CDTF (ISO 8601)
Modification time	dct:modified	0..1	Typed literal	W3CDTF (ISO 8601)
Publisher	**dc:publisher**	1..*	Instance	foaf:Organization
Creator	dc:creator	0..*	Instance	foaf:Agent
Rights	dc:rights	0..*	Literal	Non-empty string
NoIndex	ts:noindex	0..1	Typed literal	Boolean value
Genre	**ts:genre**	1	Concept	Genre Classification
Presentation type	**dc:type**	1	Concept	DCMI Type vocabulary
Format	**dc:format**	1	Typed literal	IANA MIME types
Subject	**dc:subject**	1..*	Concept	Subject Vocabularies
Keyword	ts:keyword	0..*	Literal	Non-empty string
Audience	**dct:audience**	1	Concept	Audience Classification

general public and typical tasks and goals related to personal health, so a faceted navigation system directly based on the vocabularies would likely not be very useful for users of a portal devoted to personal health issues. Instead, we constructed new, citizen-centric facets and mapped these to the underlying ontologies. The end-user facets were created using a user-centered card sorting method [33]. In the search facets, the *Subject* metadata field is represented by the four independent facets *Topic*, *Life event*, *Group of people* and *Body part*. The additional facets *Document type (genre)*, *Publisher* and *Audience* were created by using the values of the corresponding metadata fields directly.

5.1. *User Interface Design and Implementation*

The user interface of the HealthFinland portal was designed using a design process inspired by the *discount usability engineering* philosophy [24,21] that stresses the application of lightweight usability evaluations performed between incremental design updates. The design was based on experiences with earlier portals and search systems, lightweight design methods such as paper prototypes, and evaluation using heuristics and opportunistic user testing. Our evaluation was focused on a qualitative approach as we tried to find problems early on and thus guide the design process.

The first user interface sketches were drawn on paper and we made extensive use of design pattern collections such as [35] and earlier faceted search systems. The presentation of search results was designed to show rich metadata that helps users quickly find relevant items and filter out non-relevant results. The design of the search results listing was based on research by Crystal and Greenberg that evaluated the usefulness of metadata for health information seekers [3]. The display of facets and search results is shown in Figure 5.

In addition to faceted search, the portal provides recommendation 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

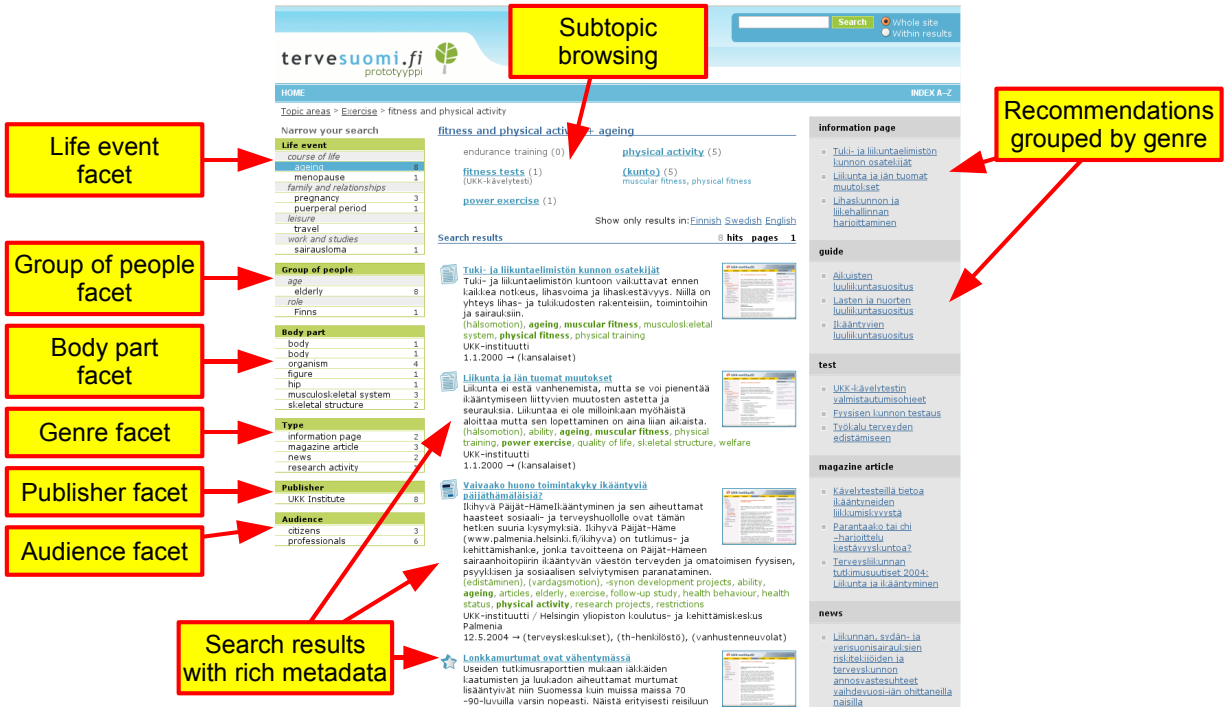

Fig. 5. HealthFinland portal with facets and search results.

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* (Figure 6), i.e., an AJAX-based semantic Web widget that can, with a small amount of JavaScript code, be incorporated into other portals to display related content items from the HealthFinland system. This way, the portal contents are exported for machine processing and mash-up applications. The implementation is similar to the related museum item display described in an earlier publication [42].

The portal was implemented as a Java Servlet application running on Apache Tomcat. It was built using the Tapestry framework and uses Jena for RDF functionality. Search and recommendation functionality was implemented using the Lucene search engine, which we enhanced to handle category and concept queries.

5.2. *Evaluation*

The portal application was 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 citizens with varying backgrounds such as students of different disciplines, administrative staff at the Helsinki University of Technology and a freelance journalist. 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. The tasks were expressed as realistic situations where the users had to solve problems or seek advice related to their own personal health. For a task to be considered successfully completed, the user was required to come up with a solution backed by information retrieved us-

Fig. 6. AJAX-based semantic Web Widget showing dynamic context-based links to the HEALTHFINLAND content. The widget has been integrated into another portal by adding a small amount of JavaScript code into the HTML page.

ing the portal and to express being satisfied with the solution. 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. Users started each task from the portal front page that shows the top level of the topic hierarchies, constructed using a user-centered card sorting process [33]. The vast majority of the first actions of test subjects took them closer to their goal, indicating that the front page design was succesful and thus that the topic structure was not confusing or misleading at this level. 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 was useful for the users who tried it.

A post-test questionnaire of subjective usability was used to calculate a System Usability Scale (SUS) [1] score. The scale represents the subjective usability of a system on a scale of 0 (worst) to 100 (best) points. The calculated SUS score of the portal was 70.8 points, representing the finding that most users considered the usability of the portal to be generally good.

5.3. *Refined User Interface*

Based on the user tests and feedback from content publishers, we made several changes to the user interface. The most significant change was the separation of search and browsing facilities, which had been combined into one interface in the prototype portal. In the new portal, both the browsing and search user interfaces are based on the categorization of documents by *Section* (audience), *Topic*, *Situation*, *Genre* and *Publisher*. Compared with the prototype portal, we have dropped the facets for *Group of people*, *Body part* and *Life event*, which were not considered very useful in our usability tests, and added

the *Situation* facet, which partly serves the same purpose as the *Life event* facet but is specific to particular topics.

The section (audience), situations, genre and publisher for a document are directly described in the metadata. The topic of a document is indirectly inferred based on a mapping between the citizen-centered portal topics and subject concepts described in the Health Promotion Ontology, as in the prototype portal.

The portal has been split into top-level *Sections* that correspond to the target audiences defined in the Audience Classification. Within each section, the content is organized into browsable *Topics*, such as the Exercise topic shown in Figure 7, and each topic may contain content areas devoted to specific *Situations*. The browsing interface resembles many traditional health portals such as BBC Health [12] and Yahoo! Health [13], but the interface is dynamically generated based on the data model that defines the sections, topics and situations which are referenced in the metadata.

The search facility is shown in Figure 8. Searches are usually initiated by entering keywords, but can be further constrained by the five available search facets.

6. Related Work

The idea of the HealthFinland portal architecture and ontology-based faceted search was inspired by earlier semantic portals such as SWED [28], the MultimediaN E-culture demonstrator [31] and MuseumFinland [15]. An early approach of creating and using an ontology infrastructure is presented in [22]. The infrastructure underlying HealthFinland was developed within the FinnONTO project [19] where HealthFinland has been a major application case, in addition to the analogous CultureSampo system [14] for cultural heritage content. Both systems are based on distributed semantic content creation, centralized ontology library services, and a portal supporting semantic search and recommendation [11].

Vocabulary and ontology development in the health domain has a long tradition and has produced very large and comprehensive systems such as MeSH, UMLS [14] and SNOMED CT [15]. In our work, existing vocabularies are used with additional mutual ontology mappings [7] and mappings to a general vocabulary (YSO).

The pitfalls of metadata about web resources have been well documented [8]. We have tried to avoid some of the problems by creating tools that make vocabularies easily available for annotators and to report problems back to the source. Transforming metadata from literal sources and databases has been discussed e.g. in [25]. In our case manual annotation using the online metadata editor SAHA together with ontology services is also used. While we haven't performed rigorous evaluations of different annotation interfaces, our own experiences are consistent with a recent end-user study of manual metadata creation using ontology services [12].

Faceted search, also called view-based search or dynamic taxonmies in IR research, has been studied by various authors for a long time [27,9,30]. Several semantic portals are based on the paradigm, both domain-specific systems such as the SWED environment portal [28] as well as generic search systems, e.g., /facet [10]. Separating end-user facets from annotation ontologies is discussed in [13]. The automatic construction of a browsing interface based on faceted metadata has been discussed in [2].

Several health portals share important aspects with our portal. The HONsearch system [16], published by the Health on the Net Foundation, allows the searching of health web sites accredited with the HONcode quality seal using keywords as well as target audiences to restrict the search. The commercial Healia search engine [17] similarly allows searches for health websites to be restricted by gender, age and heritage.

7. Conclusions

The HealthFinland 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

[12] http://www.bbc.co.uk/health/
[13] http://health.yahoo.com
[14] http://www.nlm.nih.gov/research/umls/
[15] http://www.ihtsdo.org/snomed-ct/
[16] http://www.hon.ch/HONsearch/Patients/
[17] http://www.healia.com

Fig. 7. Browsing interface that is dynamically constructed from faceted metadata.

end-user perspective, the underlying semantic technology enables a citizen-centered 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 [18] . The HealthFinland system was awarded the third prize at the Open Track of the Semantic Web Challenge 2008 competition [19] .

The National Institute for Health and Welfare is building a production version of the HealthFinland portal. The first public version of the portal [20] was published in May 2009. 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.

8. Future Work

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

Creating and maintaining the current portal, the content production infrastructure, vocabularies and metadata, requires a relatively large amount of manual work. While vocabulary maintenance is a necessary burden for most semantic applications, we are developing tools and processes that facilitate collaborative ontology development and will apply these to the maintenance of the vocabularies used

[18] http://demo.seco.tkk.fi/tervesuomi/
[19] http://challenge.semanticweb.org
[20] http://www.tervesuomi.fi

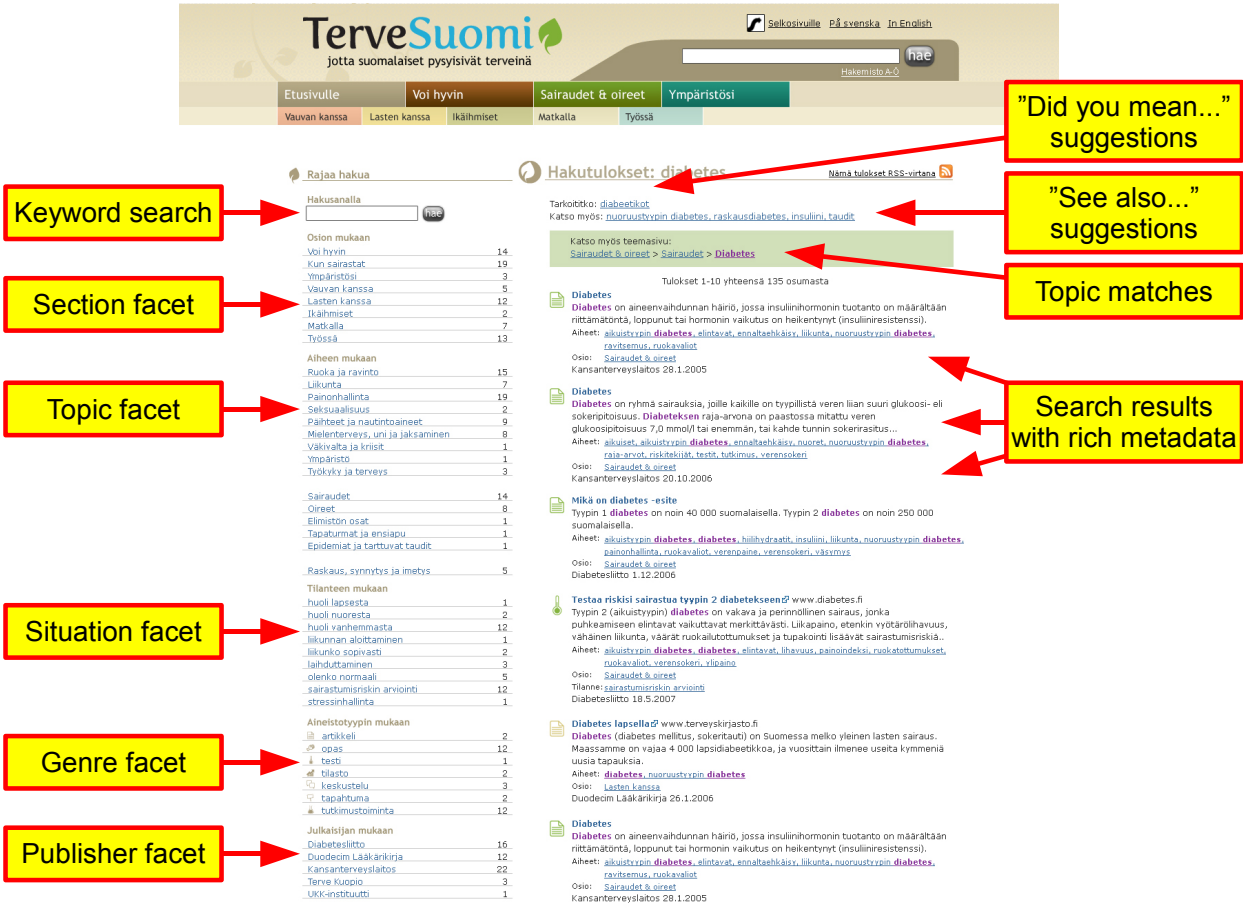

Fig. 8. Search interface, allowing queries to be restricted by five facets as well as keywords.

in HealthFinland.

Manual annotation work could further be reduced using semiautomatic annotation tools [39] to suggest, e.g., document languages and subjects based on text analysis. We are currently investigating the use of automatic categorization and topic extraction for dynamic content – such as news feeds and discussion forum postings – that could then be integrated as content sources into the portal with little or no manual effort.

The *situations* referenced by documents in the portal are currently only defined within the content repository of the portal. In the future they may be published as a controlled vocabulary on the ONKI server so that they can be referenced by content producers and third parties.

While some HealthFinland content publishers already expose their metadata, mainly through HTML `meta` tags, the aggregated and validated RDF data that the portal relies on is not currently republished back to the Web. In the future the portal will be configured to expose at least some of its metadata. We are also investigating ways to expose the portal search results in machine-processable formats such as RSS feeds, in addition to the AJAX floatlets that we already provide.

Acknowledgements

This work is a part of the national semantic web ontology project FinnONTO [21] 2003–2010, funded by the National Funding Agency for Technology Innovation (Tekes) and a consortium of 38 companies and public organizations, and the Ministry of Social Affairs and Health. The HealthFinland project is co-ordinated by the National Institute for Health and Welfare in Finland. We thank Johanna Eerola, Marko Ruohoranta, Matti Sarjakoski, Katri Seppälä

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

and Jouni Tuominen for their input to the work reported in this paper.

References

[1] J. Brooke, SUS: a 'quick and dirty' usability scale, in: Usability Evaluation in Industry, Taylor & Francis, London, 1996, pp. 189–194.

[2] A. Crystal, Facets are fundamental: Rethinking information architecture frameworks, Technical Communication 54 (1) (2007) 15–27.

[3] A. Crystal, J. Greenberg, Relevance criteria identified by health information users during web searches, Journal of the American Society for Information Science and Technology 57 (10) (2006) 1368–1382.

[4] Dublin Core Workgroup, Dublin core metadata element set, v. 1.1, http://dublincore.org/documents/dces/ (2008).

[5] Dublin Core Workgroup, Expressing Dublin Core metadata using HTML/XHTML meta and link elements, http://dublincore.org/documents/dc-html/ (2008).

[6] Dublin Core Workgroup, Expressing Dublin Core metadata using the Resource Description Framework (RDF), http://dublincore.org/documents/dc-rdf/ (2008).

[7] J. Euzenat, P. Shvaiko, Ontology matching, Springer–Verlag, Berlin, 2007.

[8] D. Hawking, J. Zobel, Does topic metadata help with web search?, Journal of the American Society for Information Science and Technology 58 (5) (2007) 613–628.

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

[10] M. Hildebrand, J. van Ossenbruggen, L. Hardman, /facet: A browser for heterogeneous Semantic Web repositories., in: International Semantic Web Conference, 2006.

[11] M. Hildebrand, J. van Ossenbruggen, L. Hardman, An analysis of search-based user interaction on the Semantic Web, Tech. Rep. INS-E0706, CWI, Amsterdam (2007).

[12] M. Hildebrand, J. van Ossenbruggen, L. Hardman, G. Jacobs, Supporting subject matter annotation using heterogeneous thesauri, a user study in web data reuse., Tech. Rep. INS-E0902, CWI, Amsterdam (2009).

[13] M. Holi, E. Hyvönen, Fuzzy view-based semantic search, in: Proceedings of ASWC 2006, Beijing, China, Springer–Verlag, Berlin, 2006.

[14] E. Hyvönen, E. Mäkelä, T. Kauppinen, O. Alm, J. Kurki, T. Ruotsalo, K. Seppälä, J. Takala, K. Puputti, H. Kuittinen, K. Viljanen, J. Tuominen, T. Palonen, M. Frosterus, R. Sinkkilä, P. Paakkarinen, J. Laitio, K. Nyberg, CultureSampo—Finnish culture on the Semantic Web 2.0. Thematic perspectives for the end-user., in: Proceedings of Museums and the Web 2009 (MW2009), Archives & Museum Informatics, Toronto, Canada, 2009.

[15] E. Hyvönen, E. Mäkela, M. Salminen, A. Valo, K. Viljanen, S. Saarela, M. Junnila, S. Kettula, MuseumFinland – Finnish museums on the Semantic Web, Journal of Web Semantics 3 (2) (2005) 224–241.

[16] E. Hyvönen, S. Saarela, K. Viljanen, Application of ontology techniques to view-based semantic search and browsing, in: Proceedings of ESWS 2004, Heraklion, Greece, Springer–Verlag, Berlin, 2004.

[17] E. Hyvönen, K. Viljanen, O. Suominen, HealthFinland—Finnish health information on the Semantic Web, in: Proceedings of ISWC 2007 + ASWC 2007, Busan, Korea, Springer–Verlag, Berlin, 2007.

[18] E. Hyvönen, K. Viljanen, O. Suominen, E. Hukka, HealthFinland – publishing health promotion information on the Semantic Web, in: Proceedings of DrMED 2008: International Workshop on Describing Medical Web Resources. The 21st International Congress on the European Federation for Medical Informatics (MIE 2008), Göteborg, Sweden, 2008.

[19] E. Hyvönen, K. Viljanen, J. Tuominen, K. Seppälä, Building a national Semantic Web ontology and ontology service infrastructure—the FinnONTO approach, in: Proceedings of ESWC 2008, Tenerife, Spain, Springer–Verlag, Berlin, 2008.

[20] JHS workgroup, JHS 143: Asiakirjojen kuvailun ja hallinnan metatiedot, http://www.jhs-suositukset.fi/suomi/jhs143 (2004).

[21] S. Krug, Don't Make Me Think : A Common Sense Approach to Web Usability, 2nd ed., New Riders Press, 2005.

[22] A. Maedche, B. Motik, L. Stojanomic, R. Studer, R. Volz, An infrastructure for searching, reusing and evolving distributed ontologies, in: Proceedings of WWW 2003, Budapest, Hungary, ACM Press, New York, 2003.

[23] E. Mäkelä, K. Viljanen, O. Alm, J. Tuominen, O. Valkeapää, T. Kauppinen, J. Kurki, R. Sinkkilä, T. Känsälä, R. Lindroos, O. Suominen, T. Ruotsalo, E. Hyvönen, in: Proceedings of First Industrial Results of Semantic Technologies, co-located with ISWC 2007 + ASWC 2007, Busan, Korea, 2007, cEUR Workshop Proc., Vol 293.

[24] J. Nielsen, Guerrilla HCI: using discount usability engineering to penetrate the intimidation barrier, in: Cost-justifying usability, Academic Press, Inc., Orlando, FL, USA, 1994, pp. 245–272.

[25] B. Omelayenko, Porting cultural repositories to the semantic web, in: Proceedings of the 1st Workshop on Semantic Interoperability in the European Digital Library (SIEDL-2008), 2008.

[26] Pew Internet & American Life Project, Online Health Search (2006).

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

[28] D. Reynolds, P. Shabajee, S. Cayzer, Semantic Information Portals, in: Proceedings of WWW 2004, Alternate track papers & posters, ACM Press, New York, New York, 2004.

[29] M. A. Rosso, User-based identification of web genres, Journal of the American Society for Information Science and Technology 59 (7) (2008) 1053–1072.

[30] m. schraefel, M. Karam, S. Zhao, mSpace: interaction design for user-determined, adaptable domain exploration in hypermedia, in: Proceedings of AH 2003: Workshop on Adaptive Hypermedia and Adaptive Web Based Systems, 2003.

[31] A. Schreiber, A. Amin, M. van Assem, V. de Boer, L. Hardman, M. Hildebrand, L. Hollink, Z. Huang, J. Kersen, M. Niet, B. Omelayenko, J. Ossenbruggen, R. Siebes, J. Taekema, J. Wielemaker, B. Wielinga, MultimediaN E-Culture demonstrator, in: Proceedings of ISWC 2006, Athens, GA, Springer–Verlag, Berlin, 2006.

[32] O. Suominen, Käyttäjäkeskeinen moninäkymähaku semanttisessa portaalissa (user-centric faceted search in a semantic portal), Master's thesis, University of Helsinki, Dept. of Computer Science (February 2008).

[33] O. Suominen, K. Viljanen, E. Hyvönen, User-centric faceted search for semantic portals, in: Proceedings of ESWC 2007, Innsbruck, Austria, Springer–Verlag, Berlin, 2007.

[34] O. Suominen, K. Viljanen, E. Hyvönen, TerveSuomi-portaalin metatietomäärittely (HealthFinland portal metadata specification), v. 2.0, Tech. rep., Semantic Computing Research Group, SW 2.0 project (2009).

[35] J. Tidwell, Designing Interfaces, O'Reilly & Assoc., Sebastopol, California, 2005.

[36] J. Tuominen, M. Frosterus, K. Viljanen, E. Hyvönen, ONKI SKOS server for publishing and utilizing SKOS vocabularies and ontologies as services, in: Proceedings of ESWC 2009, Heraklion, Greece, Springer–Verlag, Berlin, 2009.

[37] O. Valkeapää, O. Alm, E. Hyvönen, Efficient content creation on the semantic web using metadata schemas with domain ontology services, in: Proceedings of ESWC 2007, Innsbruck, Austria, Springer–Verlag, Berlin, 2007.

[38] M. van Assem, V. Malaise, A. Miles, G. Schreiber, A method to convert thesauri to SKOS, in: Proceedings of ESWC 2006, Budva, Montenegro, Lecture Notes in Computer Science, 2006.

[39] A. Vehviläinen, E. Hyvönen, O. Alm, A semi-automatic semantic annotation and authoring tool for a library help desk service, in: Emerging Technologies for Semantic Work Environments: Techniques, Methods, and Applications, IGI Group, Hershey, USA, 2008.

[40] K. Viljanen, J. Tuominen, E. Hyvönen, Publishing and using ontologies as mash-up services, in: Proceedings of the 4th Workshop on Scripting for the Semantic Web (SFSW2008) at ESWC 2008, 2008.

[41] K. Viljanen, J. Tuominen, E. Hyvönen, Ontology libraries for production use: The Finnish ontology library service ONKI, in: Proceedings of ESWC 2009, Heraklion, Greece, Springer–Verlag, Berlin, 2009.

[42] K. Viljanen, J. Tuominen, T. Känsälä, E. Hyvönen, 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.

[43] W3C, Gleaning resource descriptions from dialects of languages (GRDDL), W3C Recommendation, 11 Sept 2007, http://www.w3.org/TR/2007/REC-grddl-20070911/ (2007).