User-Centric Faceted Search for Semantic Portals

Osma Suominen, Kim Viljanen, and Eero Hyvönen

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. Many semantic portals use faceted browsing, where the facets are based on the underlying indexing ontologies of the content. However, in many cases, like in medical applications, the ontologies may be very large and complex, and do not provide the end-user with intuitive facet hierarchies for conceptualizing the content, for formulating queries, and for classifying the search results. We argue that in such cases end-user facets should be separated from the annotation ontologies, and show how to generalize the semantic view-based search paradigm to take into account this fact. A user-centric card sorting method is proposed for designing intuitive views for the end-users and a method for mapping its facets onto the indexing ontologies and search items is presented. The system has been implemented in a prototype of the semantic portal TerveSuomi.fi, a national health promotion portal in Finland.

1 Introduction

Faceted search (i.e., faceted browsing and view-based search) [1,2,3], is a search paradigm developed originally in the field of information retrieval. The idea of the scheme is to analyze and index search items along multiple orthogonal taxonomies that are called subject *facets* or *views*. From the end-users viewpoint searching is then reduced to the selections of categories along the facets. This idea has later been developed into *semantic faceted search* [3,4,5], where the facets are based on ontological structures, such as subclass and part-of hierarchies. The usefulness of faceted search has been demonstrated in many applications [5,6] and the first commercial products are already on the market[1].

A limitation of semantic faceted search is that facets based on indexing ontologies do not always provide the end-user with natural categorizations of the content for formulating queries or for organizing search result lists. In many domains, very large and complex ontologies are used for indexing by domain professionals. The point of view of indexing may differ substantially from the point of view of the end-user, who also may not be familiar with the professional terminology. The ontologies may also be too general or too specific for her needs.

[1] http://www.siderean.com/, http://www.express.ebay.com/, http://endeca.com/

E. Franconi, M. Kifer, and W. May (Eds.): ESWC 2007, LNCS 4519, pp. 356–370, 2007.

The main hypotheses underlying this paper is that end-users, such as ordinary citizens, often conceptualize the domain of discourse in terms of categories that are different from the ontological representations used by the domain specialist and content indexers. To bridge the semantic gap between end-users and professionals, we need 1) a method for finding out the end-user facets and search categorizations about the domain, and 2) a method for mapping the facet categories onto the content indexed along ontologies. To solve the facet creation problem, we present a user-centric card sorting method [7,8,9] for creating intuitive search facets for the end-users, and present results of applying the method in creating the facets for the prototype of the national semantic health promotion portal TerveSuomi.fi [10,11] in Finland. To address the mapping problem, we show how the user-centric facets can be mapped onto the ontologies used for describing the content, and be used for answering queries.

The paper is organized as follows. We first present a general architecture and a method for answering faceted queries based on user-centric facets and indexing ontologies. After this the method is applied to our case study by describing ontological metadata and the indexing ontologies which are used to describe the content. Based on this, card sorting methods for creating user-centric facets are discussed and applied for the TerveSuomi.fi portal, and a prototype implementation of the system is presented. Finally, contributions of the work are summarized, related work discussed, and future research suggested.

2 Extending Faceted Search Architechure

In semantic faceted search the documents are annotated along different facets that typically correspond to the elements (fields) of the metadata (annotation) schema used. For example, in MuseumFinland [5] the collection artifact metadata schema has nine resource-valued properties such as *Artifact type* and *Material* whose values are taken from a set of seven orthogonal *indexing ontologies*. A single ontology, such as "Places", can be used for expressing values of several different elements, such as "Place of Manufacture" and "Place of Usage". When the same place is selected in the "Place of Manufacture" or "Place of Usage" facet for querying, different result sets are obtained. Although facets share hierarchical structures of the indexing ontologies, the facet categories are different from the corresponding ontology concepts and have different facet-wise URIs. The resulting facet hierarchies will be called *faceted ontologies*.

By using a set of logical projection rules each facet-URI can be associates with a set of related search items [12]. The *extension* $ext(S)$ of a facet-URI S is the set of all search items associated with it and any of its subcategories. The faceted search query in semantic faceted search with n facets is a conjuctive Boolean expression $Q = S_1 \wedge ... \wedge S_n$ where each S_i is a Boolean expression of the facet ontology category (URIs). The result set is obtained by interpreting the logical operations of Q as set operations over the search item set E in the following way: $S \wedge T \equiv ext(S) \cap ext(T)$, $S \vee T \equiv ext(S) \cup ext(T)$, $S - T \equiv S \wedge \neg T \equiv ext(S) - ext(T)$, and $\neg S \equiv E - ext(S)$.

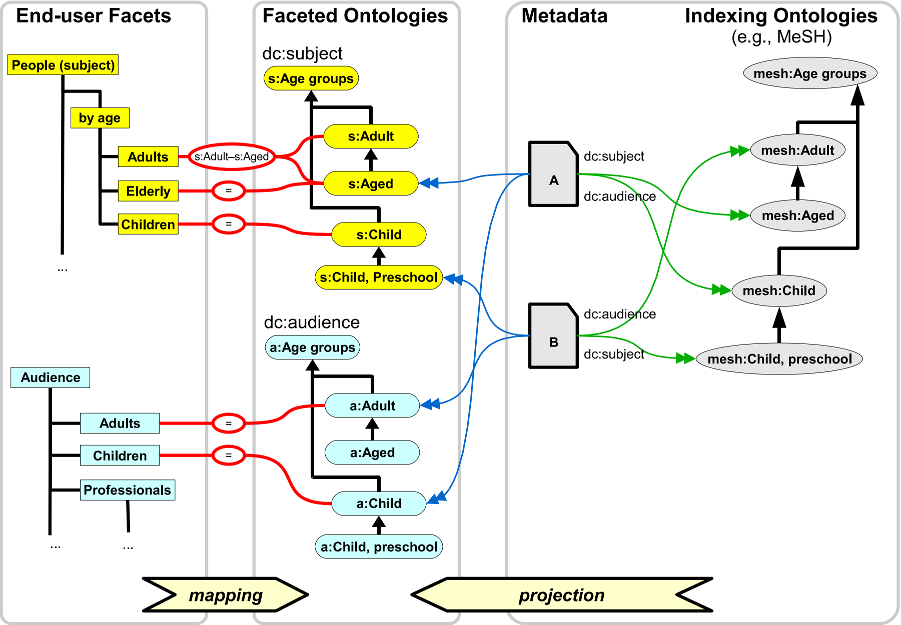


Fig. 1. Components of an end-user-centric view-based semantic search framework

In MuseumFinland, faceted ontologies are used directly as querying facets (with some filtering). In the case of TerveSuomi.fi this is not feasible but a set of m user-centric facets is created for formulating the query $Q_U = U_1 \wedge \ldots \wedge U_m$, where U_i is a category in a user-centric facet i. In order to map such queries onto queries at the faceted ontology level, each user-centric facet category can be defined as a Boolean expression of the faceted ontology categories. This means that queries expressed in terms of end-user facet categories can be reduced into faceted ontology queries that can be processed with a semantic faceted search engine such as Ontogator [13].

Figure 1 depicts how the search documents can be found using user-centric facets, indexing facets and indexing ontologies in our view-based semantic search scheme. The documents are indexed with ontologies using resource-valued metadata fields `dc:subject` and `dc:audience` whose values are taken from the indexing ontology (e.g., MeSH). The document A tells about (dc:subject) aged people (elderly people in layman terms) and is intended for children (dc:audience). Similarly, the document B is about preschool children and is intended for adults. The corresponding two faceted ontologies, projected from the indexing ontologies, are seen in the middle left in different namespaces s (subject) and a (audience). Some of the extension of the categories are: ext(s:Aged)=ext(s:Adult)={A}, ext(s:Child)={B}, ext(a:Child)={A} and ext(a:Adult)=ext(a:AgeGroup)={B}. The extensions can be projected by simple logic rules based on the metadata. If faceted ontologies were used for searching (like in MuseumFinland), then the selection *s:Adult* would return the document A and *a:Adult* would return the document B. By introducing end-user facets, the queries can be expressed in terms of new facet categories. For example, the end-user facet category *Adults* is defined as $Adults \equiv ext(s : Adult) - ext(s : Aged)$, and the document A would not be returned with the facet selection *Adults*, because the document is about aged people which has been excluded from the end-user facet category *Adults*.

This model can be extended to deal with uncertainty and relevance by using the "fuzzy view-based search" approach presented in [10] in the following way. Fuzzy annotations can be used if exact classifications are not appropriate. For example, the boundary between children and adults is fuzzy. Therefore we could say that the dc:audience of the document B in figure 1 is not the crisp set $\{mesh : Adult\}$ but rather the fuzzy set $\{(mesh : Adult, 0.8), (mesh : Child, 0.2)\}$ indicating that the document is targeted to some degree also to children. In [11] a method for determining such fuzzy annotations based on ontologies and the *tf-idf* method is presented. In this way the extensions of the indexing facets can be seen as fuzzy sets where the membership values are based on fuzzy annotations and are interpreted as a measure of relevance. By defining the Boolean operators used in faceted search as fuzzy Boolean operators, the relevance of hits in the search results can be determined—an important feature missing in the traditional faceted search paradigm. In the example above, the document B would be less relevant when looking for material targeted to children than to adults. It is also possible to generalize the mappings into fuzzy mappings by attaching a membership function value to them, indicating only partial match between an end-user category and its definition in terms of the facet ontology categories, and by interpreting the mapping as a fuzzy inclusion.

3 Ontological Metadata for a Health Promotion Portal

When building a semantic portal, one key decision is to choose which ontologies are used for indexing the content and whether existing ontologies can be used compared to building custom ontologies. Typically suitable ontologies exist, which have been created for a similar domain but a different purpose. In such cases the decision has to be made between using them as-is, modifying them to suit the purposes of the portal, or creating a new ontology from scratch.

Using an existing, established ontology has several advantages. Creating a new ontology requires substantial amounts of manual work, which can be avoided by reusing an existing ontology. An existing and established, large ontology is also more likely to have broad and deep coverage of concepts within its domain, which will allow documents to be annotated to very specific concepts, whereas a custom-made ontology might only cover the topic areas and concepts that are relevant for the need of the semantic portal. Finally, reusing a shared ontology furthers the vision of the Semantic Web [14] by allowing semantic interoperability between different systems, as long as they use the same ontology (or a compatible one, interlinked by semantic mappings). On the other hand, existing ontologies may differ from the goals of the portal in their scope or the point of view. These and other problems, such as licensing and technical issues, involved in reusing existing ontologies have to be balanced against the benefits.

In the case of the health promotion portal TerveSuomi.fi, the information items of interest are web-accessible publications such as web pages and PDF documents, and they are described using Dublin Core[2] metadata such as `dc:subject` which contains the subject topic(s) of the given document. We decided to use the following ontologies[3] as indexing ontologies for the subject field: the Finnish General Upper Ontology (YSO)[4] [15] which is based on the General Finnish Thesaurus YSA[5] that is widely used in Finland for indexing contents of various kinds, Medical Subject Headings (MeSH)[6] and the European Multilingual Thesaurus on Health Promotion (HPMULTI)[7]. The reason for combining them (see also Table 1) was that none of them was alone adequate for describing the topics of the whole variety of content documents in the portal. YSO is broad but too general with regard to medical content. On the other hand, MeSH is too focused on clinical healthcare while HPMULTI has very narrow coverage, focusing exclusively on health promotion terminology.

Table 1. Core subject ontologies of the TerveSuomi.fi portal

Name	YSO	MeSH	HPMULTI
Publisher	National Library of Finland & FinnONTO	National Library of Medicine, USA	European Commission
# concepts	23 000	23 000	1 200
Languages	Finnish; Swedish and English under construction	English; Finnish and Swedish translations available	Multilingual, including Finnish, Swedish and English
Intended use	Cataloging of material published in Finland	Cataloging of biomedical documents	Cataloging of material on health promotion
Intended user group	Librarians	Medical professionals; librarians within the field of medicine	Professionals involved in health promotion
Examples of concepts	Travellers Water pipes Cities Vegetables	Metabolic Syndrome X Endocrine Disruptors Biopsy, Fine-Needle DNA Damage	Traffic accidents Behavioural change Voluntary work Sunburn

To prevent the creation of internal semantically incompatible islands within the portal, corresponding concepts in each ontology had to be mapped to each other. Mapping the ontologies was done using three complementary approaches:

[2] http://dublincore.org

[3] The term *ontology* is used in a broad sense, covering also thesauri in the sense that they are formal, explicit classifications for describing the content of documents.

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

[5] http://vesa.lib.helsinki.fi

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

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

a) using available, existing mappings between MeSH and HPMULTI; b) creating automatic mappings between MeSH and YSO based on textual matching of concept labels; and c) manually mapping HPMULTI to YSO. The result was a interlinked combination ontology, where YSO provides the upper concepts and MeSH and HPMULTI the more exact concepts.

When analysing the ontologies, we noticed that these ontologies were created for use by professionals and their intended use is somewhat different from their use within the portal. This disparity between the points of view of the ontologies and the users of the portal manifests itself in many ways. First, the concept hierarchies are often inappropriate for the portal. MeSH e.g. uses deep hierarchies with complex subclassification criteria, and YSO contains many generic concepts that would probably only be confusing as facet categories. Second, concepts in professional classifications have typically been labelled using professional terminology instead of layman terms used by the end-users. On the other hand, there are many terms in everyday use that are not used by professionals due to their ambiguity. A portal must also be able to deal with queries based on such terminology even if it is not considered appropriate from a professional perspective. As an example of the problems involved, consider the MeSH top-level categories *Anthropology, Education, Sociology and Social Phenomena*, *Biological Sciences* and *Technology and Food and Beverages*. Such categories in a facet would not be very good starting points for a person looking for information about dieting. Another user might not realize that information about breast cancer can be found under the MeSH concept *Breast Neoplasms*.

4 Creating User-Centric Search Facets with Card Sorting

To solve the mismatch between the indexing ontologies and the expectations of the end-users, we propose using user-centric design practices to construct a custom classification system for the portal based on the users' expectations and their mental models, with the intent of later mapping it to the underlying ontological concepts to provide semantically sound faceted browsing and other functionalities.

A practical method for gathering information about the end-user's mental models of an information space, i.e., how users of a website tacitly group, sort and label tasks and content, is the *card sorthing* method [7,8,9]. A card sorting study is typically performed using index cards, with each card bearing the title and possibly a short description of an individual document. The study is then performed on volunteers that are asked to sort the cards into piles based on intuitive feeling of similarity or relateness of the given cards, and to give the piles descriptive names. This variation of card sorting where the categories (piles) are not given beforehand but are created by the participants is called *open card sorting* [9,16]. Card sorting doesn't directly give the designer a finished categorisation structure, but provides insight into the design choices for creating such a structure.

4.1 Selecting Card Contents

When using card sorting for creating user-centric facets for organizing ontologically indexed content, we propose using the ontological concepts as values of the index cards. To avoid overwhelming the participants of the study, only the most frequently used indexing concepts (based on a sample of indexed content) excluding overly general concepts should be used in the cards.

In our case, we created a list of all the concepts that occurred in our annotated content items (n=523). These 1722 concepts were then ranked by their frequency of occurrence. Concepts with only a few documents were pruned, as well as overly general concepts such as *health* and *health promotion*. Finally, concepts judged to be uninteresting or unnecessary from the point of view of the study were eliminated. These included, e.g., geographical locations, individual organizations, abstract concepts such as *Development* and *Promotion* as well as concepts that were considered very similar to others that were already on the list[8]. The pruning brought down the number of concepts to 177, which was deemed acceptable for the card sorting exercise. The labels of the concepts were printed on index cards together with numeric identifiers for ease of analysis.

4.2 Performing the Card Sorting

The study participants should be representative of the expected users of the system. Nielsen recommends using 15 participants [17] while Maurer & Warfel recommends seven to ten individuals [16]. Each participant is advised to group the cards into piles according to their meaning or topical similarity. Participants are asked to think aloud, especially when facing difficult decisions. During the exercise, the facilitator takes notes of important events and insightful comments made by participants during the experiment. If a pile becomes very large, the participants are instructed to split it into smaller parts. After sorting the cards into piles, they are asked to write down a descriptive label for each pile.

In our case, the card sorting study was performed on volunteers that were chosen to represent potential users of the system. A total of ten individuals of varying ages and backgrounds participated, with three of them doing the exercise as a group while the others performed the study alone. Thus, a total of eight rounds were performed. The raw data obtained during the card sorting study is a set of labeled piles of cards such as those shown in Figure 2.

4.3 Creating the Result Categories Based on the Card Piles

When analyzing card sort results, both qualitative and quantitative aspects need to be considered [9]. When card sorting is used for getting input into the design of website navigation, gaining insight from the data is of foremost importance; whether that requires a rigorous statistical analysis depends on the situation at hand. In many projects, simply “eyeballing” the data may provide enough

[8] E.g., only a sample of the dozens of food items such as *Meat* and *Cheese* were kept.

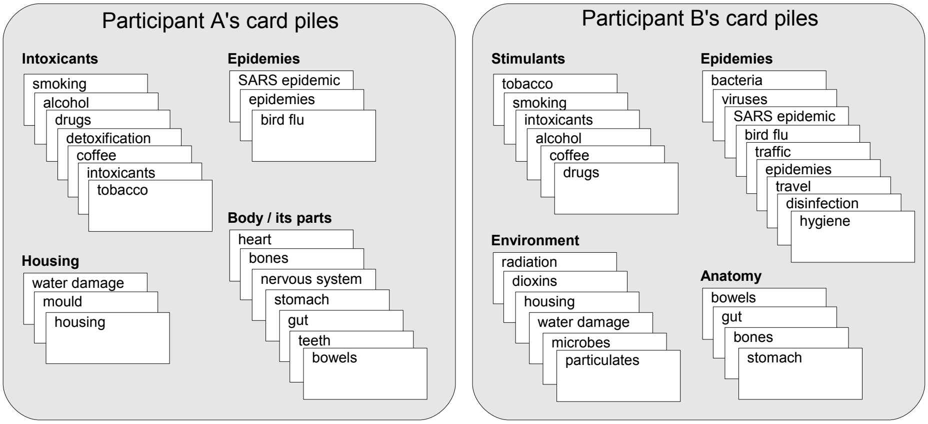


Fig. 2. Examples of card piles created by study participants

insight to create a workable design [8]. In our case, we used a spreadsheet template [18] to calculate some metrics such as card co-occurrence and the average number of cards in each category, but did not perform a full-fledged statistical cluster analysis. However, automatic tools have been created that perform cluster analysis and create tree diagrams that might be used directly as a basis for constructing web site navigation [19].

To create the result categories, we used the following processing steps: First, the categories created by individual participants were manually clustered to create a standardized set of categories for the purposes of analysis. As an example of the clustering process, the category *Body and its parts* created by one participant was considered the same as category *Anatomy* created by another participant, and these were both mapped to the standard category *Body part*. The clustering resulted in 29 standard categories and a mapping of each participant's categories to these. Sometimes similar categories created by a single participant were mapped to the same standard category for the purposes of analysis (e.g., *Body and its parts* and *Teeth* were both mapped to *Body part*), and not all standard categories were present in all user categories (e.g., not all participants had included a category for *Weight control*).

The second step of analysis was to enter the raw data about user-created categories and their contents (individual concepts) into the spreadsheet, using the above defined mappings.

The third step was to actually analyze the data, looking for patterns of interest. The analysis spreadsheet revealed, for example, that there was a high agreement about the existence of a category for body parts (all participants had included such a category) as well as the contents of that category. E.g., all participants had placed the concepts *Stomach* and *Skeletal system* in that category. On the other hand, while three participants had included a category for well-being, there was low agreement about the contents of that category. The interpretation for these results was that participants (and, by extension, users of the portal) have a clear mental model of body parts as a category distinct from, say, food and nutrition issues, while a category for lifestyle doesn't invoke such a clear notion of distinctness. Based on the analysis, we were able to pick good candidates for top-level facets as well as construct part of their contents. The analysis revealed that *Body part*, *Group of people* and *Life event* were popular categories. Perhaps more importantly, they were at least somewhat orthogonal towards each other and the rest of the categories, so they were chosen to be presented as separate facets. The rest of the categories were then used to create a fourth facet called *Topic*.

For each of the remaining categories, we had to decide whether to a) discard the category altogether (in cases of low agreement), b) use it as a top-level category or c) place it below a top-level category in the hierarchy. The hierarchical relations between categories could not directly be seen from the card sorting analysis. However, hints about these could be found in the notes made during the card sorting sessions, e.g., situations where a participant had split a large pile into smaller components.

Some of the discarded categories included *Oversensitivities*, *Disease prevention and self-help* and *Health problems*. The resulting hierarchy is quite shallow; additional levels may need to be added using other methods such as *laddering* [20]. However, for the purposes of our portal we have expanded the hierarchy simply by examining the underlying ontologies and building up the hierarchy by mirroring their structure, while trying to make sure that the terminology and groupings are suitable for end-users. We feel that while the design of lower hierarchy levels could benefit from user-centric design methods, the issues here are not as critical as the choice of facets and their topmost categories.

4.4 Finalizing and Evaluating the Categorisation

When an initial version of the facets is created using the card sorting method described above, the result should be evaluated and possibly reviewed both by additional user testing and by domain experts. One way for evaluating the result categories with users is to do new rounds of card sorting using the closed card sort method where the categories are given beforehand and the study participants are asked to sort the cards into those piles. The intuitiveness of the categories can be estimated based on how well the results of the closed card sorting matches the initial facets.

In our case, the user testing of the *Topic* facet (see Figure 3) using the closed card sorting method was done with two volunteers who were asked to sort the ontological concepts into the suggested top-level categories. The results were encouraging: participants placed nearly all concepts in the category that was intended by the designer. More user evaluations would probably need to be done to find subtler errors.

Additionally, an expert review of the initial facets was done by health promotion experts, which revealed some problems. The category *Catastrophes & Epidemies* was considered problematic: the two concepts are not very closely related and lumping them together may send false signals to users of the portal.

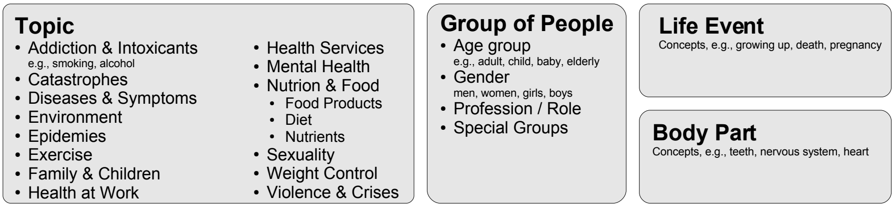


Fig. 3. The finalized end-user facets with some examples of the concepts

A new look into the analysis process revealed that these two concepts had been paired together during the clustering phase, and in fact only one user had created a category where both aspects were present – a clear mistake in the clustering. Thus, separating the two aspects into their own top-level categories was an easy decision. Another problem discovered by domain experts was that there was no category for issues related to occupational health, such as the hazards of dangerous chemicals used at work. Such concepts were not very well represented in the set used for the card sort, possibly because the initial set of documents lacked documents specific to occupational health. A new top-level category for occupational health issues was created, with subcategories taken from ontologies as well as classifications used on existing websites on the topic. The finalized facets are presented in Figure 3.

More generally, the lesson learned was that skewed initial data and errors during the analysis may cause subtle errors in the hierarchy. However, using user evaluations and expert reviews helps alleviate at least some of the problems.

5 Mapping User-Centric Facets to Ontologies

When the initial user-centric facets have been created using the method above, the facets should be logically mapped to the ontological facets as described in section 2. Since the card sorting is done using a selection of typically used ontological concepts, and since the relations between the standard categories and these concepts are known, these relations can be directly used as mappings between the facets and the facet ontologies. To make the mapping comprehensive, additional concepts are needed, which must be added manually. E.g., if the concept *Food* was used in the card sorting but the concept *Nutrition* was not, the latter might be relevant also for a facet category of *Nutrion and Food*.

We have decided to represent the facet hierarchies described in the previous chapter in RDF using the SKOS Core vocabulary[9] and their connections to the underlying ontologies using the SKOS Mapping vocabulary[10]. Each facet is described as a `skos:ConceptScheme` and each facet category is represented as a `skos:Concept` with a human-understandable `skos:prefLabel`. The facet

[9] http://www.w3.org/2004/02/skos/core/ (URI prefix `skos`)

[10] http://www.w3.org/2004/02/skos/mapping/ (URI prefix `skosmap`)

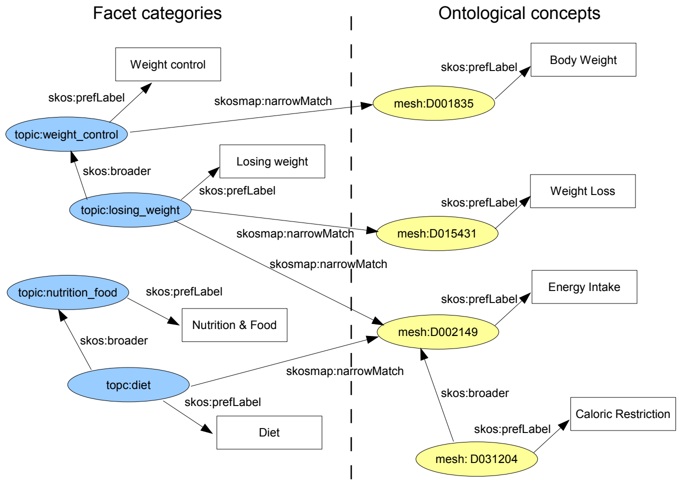


Fig. 4. Examples of mappings between facet and ontology concepts. The URI prefix `topic` refers to the *Topic* facet and `mesh` to the indexing ontology MeSH.

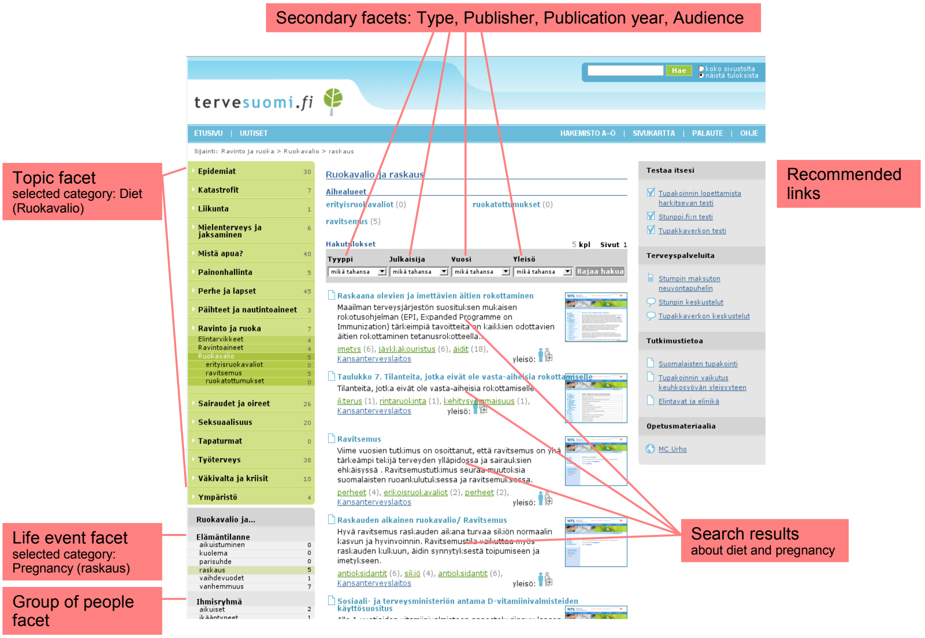
hierarchies are represented using `skos:broader` and `skos:narrower` relationships. Mappings to the underlying ontological concepts are represented using `skosmap:narrowMatch`. This mapping is a subset of the mappings described in Section 2; such more complex mappings can also be expressed using the SKOS Mapping vocabulary and will likely be used in the future. An example of mappings between facet categories and ontological concepts is shown in Figure 4.

This representation implies, by the SKOS inference rules, that a category within a facet contains (is the subject of) all documents that are annotated with one of the ontological concepts that the category has been mapped to. In addition, the category subsumes its child categories, and the ontology concept is the subject of all its narrower concepts. Thus, in Figure 4, the category `topic:weight_control` will contain all documents indexed against any of the MeSH concepts in the figure due to the `skos:broader` and `skosmap:narrowMatch` relations present.

6 Prototype Implementation

As a proof of concept, the methods discussed have been implemented in the prototype of TerveSuomi.fi[11] where the faceted search functionality has been created using the faceted search engine Ontogator [13]. Figure 5 shows the user interface, where the user has selected the category *Diet* from the *Topic* facet and the category *Pregnancy* from the *Life event facet*. The result of this faceted search query is the list of links to web pages. The user could now either visit

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



Raskaana olevien ja imettävien äitien rokottaminen
Maailman terveysjärjestön suosituksen mukaisen rokotusohjelman (EPI, Expanded Programme on Immunization) tärkeimpiä tavoitteita on kaikkien odottavien äitien rokottaminen tetanusrokotteella...
imetys (6), jäykkäkouristus (6), äidit (18),
Kansanterveyslaitos yleisö:

Taulukko 7. Tilanteita, jotka eivät ole vasta-aiheisia rokottamiselle
Tilanteita, jotka eivät ole vasta-aiheisia rokottamiselle
ikterus (1), rintaruokinta (1), kehitysvammaisuus (1),
Kansanterveyslaitos yleisö:

Ravitsemus
Viime vuosien tutkimus on osoittanut, että ravitsemus on yhä tärkeämpi tekijä terveyden ylläpidossa ja sairauksien ehkäisyssä . Ravitsemustutkimus seuraa muutoksia suomalaisten ruoankulutuksessa ja ravitsemuksessa.
perheet (4), erikoisruokavaliot (2), perheet (2),
Kansanterveyslaitos yleisö:

Raskauden aikainen ruokavalio/ Ravitsemus
Hyvä ravitsemus raskauden aikana turvaa sikiön normaalin kasvun ja hyvinvoinnin. Ravitsemustila vaikuttaa myös raskauden kulkuun, äidin synnytyksestä toipumiseen ja imetykseen.
antioksidantit (6), sikiö (4), antioksidantit (6),
Kansanterveyslaitos yleisö:

Sosiaali- ja terveysministeriön antama D-vitamiinivalmisteiden käyttösuositus

Fig. 5. TerveSuomi.fi portal user interface

some of the resulted web pages or modify the query by selecting, e.g., additional facets or by clicking on context based semantical recommend links on the right.

7 Discussion

This paper argued that card-sorting combined with mappings provides a promising approach for designing and implementing semantic view-based search based on user-centric facets.

7.1 Contributions

The main benefits of separating end-user facets from content indexing ontologies are: First, more intuitive and useful user interfaces can be provided. Second, the same ontologically annotated metadata can be re-used for different use cases and interfaces without changing the metadata or the content by defining new alternative user-centric facets and mappings. This flexibility would not be achieved if the metadata were described using application-specific or user-centric categorizations directly. For example, the same metadata could be used to create both a professional facet and a citizen's facet to the same content, where the professional facet is more directly based on the indexing ontologies and the citizen facets more on the various information needs of ordinary life.

The downside of using user-centric facets is the extra work needed in creating them and in mapping search categories onto annotation ontologies. Also, if the card sorting is based on non-representative example annotations, the resulting user-centric facets might not be optimally designed when more content is added to the portal. Therefore, readjustments to the user-centric facets might be needed based on, e.g., feedback from the users.

7.2 Related Work

In earlier semantic portals based on the faceted browsing paradigm, the facets have been automatically created from the underlying ontological hierarchies using projection rules (e.g. [12]). A distinction can be made between systems where the ontologies are created to become facets in the user interface and systems that use pre-existing general purpose ontologies. The former group includes MuseumFinland [5] and SWED[12], whereas /facet[13] [6] is an example of the latter approach. The problems of matching the hierarchical structure of the ontology with user needs and expectations only become apparent in the latter case, as the point of view of the original ontology may differ a lot from the end-users' mental models of the information space. In /facet, the automated facet generation sometimes results in a user interface that is hard to use [6].

Another approach for creating a navigational hierarchy based on an ontology is presented by Stoica & Hearst [21,22]. Their system uses the WordNet lexical ontology as a basis for creating a hierarchical classification which can then be used in faceted browsing. The Castanet algorithm simplifies the WordNet IS-A hierarchy by eliminating branches that aren't represented in the document collection as well as unnecessary levels of the hierarchy. The resulting taxonomies can be used either as-is or after some manual adjustments. However, the relationship of Stoica & Hearst's work with ontological metadata is weak: WordNet is only used as a basis for creating the navigational hierarchies, and the document metadata is later assumed to reference the newly created taxonomy directly.

Card sorting has been previously used in the construction of ontologies as a means of knowledge elicitation. While card sorting is usually performed manually outside the ontology engineering process, a computerized card sorting plugin has been developed for the Protégé[14] ontology editor [23]. However, the focus of this work is on the ontology creation process itself; there is no direct intent of using the resulting ontology in a search-oriented user interface.

7.3 Future Work

We are currently implementing a more finalized prototype of the semantic portal TerveSuomi.fi. After this, user tests should be done to evaluate the prototype and the underlying hypotheses such as the end-user-centric facets. We are currently also investigating how ontologies could be used to model health care services

[12] http://www.swed.org.uk

[13] http://slashfacet.semanticweb.org

[14] http://protege.stanford.edu

using methods presented in [24]. In the future, the portal may be extended to incorporate access to personal medical records and health care services.

Acknowledgements

We thank Markus Holi and Petri Lindgren for their input to the work reported in this paper. This work is part of the FinnONTO project[15], funded mainly by the National Funding Agency for Technology Innovation (Tekes) and the Ministry of Social Affairs and Health. The TerveSuomi.fi project is co-ordinated by the National Public Health Institute in Finland (Project Coordinator Eija Hukka).

References

1. Pollit, 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.
2. 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
3. Hyvönen, E., Saarela, S., Viljanen, K.: Application of ontology techniques to view-based semantic search and browsing. In: The Semantic Web: Research and Applications. Proc. of the 1st European Semantic Web Symposium (ESWS 2004). (2004)
4. Mäkelä, E., Hyvönen, E., Sidoroff, T.: View-based user interfaces for information retrieval on the semantic web. In: Proceedings of the ISWC-2005 Workshop End User Semantic Web Interaction. (Nov 2005)
5. Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S.: Museumfinland – finnish museums on the semantic web. Journal of Web Semantics **3**(2) (2005) 25
6. Hildebrand, M., van Ossenbruggen, J., Hardman, L.: /facet: A browser for heterogeneous semantic web repositories. In Cruz, I.F., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L., eds.: International Semantic Web Conference. Volume 4273 of Lecture Notes in Computer Science., Springer (2006) 272–285
7. Rugg, G., McGeorge, P.: The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. Expert Systems **14**(2) (1997) 80–93
8. Nielsen, J., Sano, D.: Sunweb: User interface design for sun microsystem's internal web. In: Proceedings of the 2nd World Wide Web Conference, Chicago, IL. (Oct 17-20 1994) 547–557
9. Rosenfeld, L., Morville, P.: Information Architecture for the World Wide Web. second edn. O'Reilly (2002)
10. Holi, M., Hyvönen, E.: Fuzzy view-based semantic search. In: Proceedings of the 1st Asian Semantic Web Conference (ASWC2006), Beijing, China, Springer-Verlag (September 3-7 2006)
11. Holi, M., Hyvönen, E., Lindgren, P.: Integrating tf-idf weighting with fuzzy view-based search. In: Proceedings of the ECAI Workshop on Text-Based Information Retrieval (TIR-06). (Aug 2006)

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

12. 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 (September 4-8 2006) 370–376
13. Mäkelä, E., Hyvönen, E., Saarela, S.: Ontogator — a semantic view-based search engine service for web applications. In: Proceedings of the 5th International Semantic Web Conference (ISWC 2006). (Nov 2006)
14. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **284**(5) (May 2001) 34–43
15. 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). (Nov 2005)
16. Maurer, D., Warfel, T.: Card sorting: a definitive guide. Boxes and Arrows (Apr 7 2003) http://boxesandarrows.com/S1937.
17. Nielsen, J.: Card sorting: How many users to test (Jul 19 2004) Alertbox column, http://www.useit.com/alertbox/20040719.html.
18. Lamantia, J.: Analyzing card sort results with a spreadsheet template. Boxes and Arrows (Aug 26 2003) http://boxesandarrows.com/S1708.
19. Dong, J., Martin, S., Waldo, P.: A user input and analysis tool for information architecture. In: CHI'01 extended abstracts on Human factors in computing systems. (March 31 – Apr 05 2001)
20. Rugg, G., Malcolm, E., Mahmood, A., Rehman, N., Andrews, S., Davies, S.: Eliciting information about organizational culture via laddering. Journal of Information Systems **12**(3) (2002) 215–230
21. Stoica, E., Hearst, M.: Nearly-automated metadata hierarchy creation. In: Proceedings of HLY-NAACL'04, Boston. (May 2004)
22. Stoica, E., Hearst, M.: Demonstration: Using wordnet to build hierarchical facet categories. In: Proceedings of the International ACM SIGIR Workshop on Faceted Search, Seattle, WA. (Aug 2006)
23. Wang, Y., Sure, Y., Stevens, R., Rector, A.: Knowledge elicitation plug-in for Protégé: Card sorting and laddering. In: Proceedings of the 1st Asian Semantic Web Conference, Beijing, China. (Sep 3-7 2006)
24. Laukkanen, M., Viljanen, K., Apiola, M., Lindgren, P., Hyvönen, E.: Towards ontology-based yellow page services. In: Proceedings of WWW2004 Workshop, Application Design, Development, and Implementation Issues. (May 2004)