ABSTRACT
This paper presents conceptual assumptions about the interaction between the structural specificities of a thesaurus and the quality of a thesaurus-based application output. So far hardly any literature exists that discusses thesaurus modelling requirements with respect to the following thesaurus-specific application areas: classifying, indexing, autocomplete, query expansion, recommendation and glossaries. By looking at these application areas the authors compare the structural attributes of SKOS and discuss their functional relevance. The authors conclude that taking these assumptions into account can significantly support application-oriented thesaurus modelling hence incrementally improving thesaurus-based applications in terms of modelling scope and effort. An empirical testing of these assumptions is subject to future work.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous; D.2.8 [Software Engineering]: Metrics—complexity measures, performance measures

General Terms
Theory

Keywords
thesaurus, SKOS, quality assurance, semantic web application

1. INTRODUCTION
Thesauri can be used to support various application scenarios like Autocomplete, Faceted Search & Browsing, Recommendations or Glossaries. Herein thesauri usually perform the function of harmonising terminologies, controlling vocabularies and or support the user in browsing through a concept space [10]. Despite a long research tradition in thesaurus quality assurance little attention has so far been paid to the interaction between the structural specificities of a thesaurus and the quality of output with respect to differing application scenarios supported by the thesaurus. Although several initiatives exist that focus on thesaurus and metadata quality in terms of expressivity and structural soundness ([7], [11] existing ISO standards like [1], [2] and basic thesaurus & organisation system literature i.e. [4]), these approaches do not take the envisioned application into account, thus being of limited relevance for applied thesaurus modelling.

This paper is aiming at closing this gap by taking a look at the structural specificities of thesauri and their relevance in improving the output quality of a specific application. It is based on the assumptions that

H-1 The structural attributes of a thesaurus are of varying relevance for specific application scenarios.

H-2 The modelling principle of a thesaurus has a direct effect on the quality of a thesaurus-based application.

To investigate into these problems the paper has been composed of the following sections. The next section gives an overview over related work in the domain of thesauri for web-based applications. This analyses starts off with general look at thesaurus quality criteria but then takes a specific look at the W3C recommendation SKOS\(^1\) which has been widely accepted as a reference model for thesaurus-based applications on the (semantic) web. In the next section the implications for the development of thesauri for different applications scenarios are discussed. In the conclusion we summarize our findings and relate them to the hypotheses we defined in the introduction.

2. STRUCTURAL SPECIFICITIES OF THESAURI FOR SKOS-BASED APPLICATIONS
Since its first release in 2004 the W3C recommendation SKOS (Simple Knowledge Organisation System) has been utilized by several semantic web applications as a lightweight model to support interoperability at the terminological and

\(^1\)http://www.w3.org/2004/02/skos/
Economic like Agrovoc cabularies. Transformations have been made for thesauri existing vocabularies are offering SKOS versions of their vocabulary in web-oriented formats like SKOS is that more and more sign for the importance of having controlled vocabularies. SKOS features a concept-oriented approach, with a concept being "An idea or notion; a unit of thought." (as defined in the SKOS definition) itself that can be represented with an URI. Another sign for the importance of having controlled vocabularies in web-oriented formats like SKOS is that more and more existing vocabularies are offering SKOS versions of their vocabularies. Transformations have been made for thesauri like Agrovoc, Eurovoc®, GEMET® and STW Thesaurus for Economic but also for other types of controlled vocabularies like subject headings (MeSH®, LCSH® etc.).

Despite the broad uptake of SKOS, research in the interaction between the modelling paradigm and the quality of the application output is comparably scarce. Wang et al. [12] have conducted an experiment on the precision and relevance of automatic artwork recommendations with respect to the underlying semantic properties. And recently Kless & Milton [7] have developed a measurement construct to evaluate the intrinsic quality of thesauri mainly based on the framework for information quality developed by Stvilia et al. [11] and the measurements constructs defined by Soergel [9].

In this paper we will concentrate on thesauri as the type of controlled vocabulary that offers the highest level of expressivity with a focus on a concept-oriented thesaurus model. In the following we will try to show that different application scenarios demand different structural specificities of a thesaurus.

3. ASSUMPTIONS ON STRUCTURAL ATTRIBUTES FOR APPLICATION-SPECIFIC THESAURU

According to our hypotheses, the structure of a thesaurus influences the quality of the application output. With reference to the work of Kless & Milton[7], who defined general (intrinsic) quality criteria for thesauri, we discuss the relevance of structural SKOS attributes for the application scenarios defined above. Table 1 gives an overview of the different applications areas with respect to the requirements of the structural attributes created for the application types.

Filtering / Classification.

A thesaurus can be used to filter, browse or classify content by categories. As learning curves for complex classifications are steep, a static hierarchy with a defined scope (limited number of concepts) is preferable compared to a dynamic one. Hence the quantity of valid concepts and labels is restricted by the application. Equivalence relations are relevant for categorization as they increase the semantic consistency of a thesaurus, while polyhierarchies and homonyms should be avoided as they increase complexity. The hierarchical depth is restricted by the application. Associative relations, definitions and notes are not relevant for classification purposes.

Indexing.

A thesaurus can improve standard indexing functionalities for documents (statistical or linguistic) by providing domain knowledge for the extraction resulting in better indexing results. The higher the domain specificity of a thesaurus, the better the indexing results will be. Hence the number of concepts and labels within a thesaurus is restricted by the scope of the domain. Equivalence relations are highly relevant for indexing documents as they increase the lexical explorativity of a document corpus, while the relevance of hierarchical and associative relations is not relevant for indexing purposes as they mainly play a role for retrieval of indexed content objects which is covered in the recommendation scenario (see blow). Indexing will go hand in hand with statistical and linguistic approaches for extracting terms. This can also support a semi-automatic thesaurus maintenance approach providing new terms by determining frequently extracted terms not found in the thesaurus and suggesting them as new concepts.

Autocompletion.

A thesaurus can support autocomplete functionalities, the syntactic normalization of free text input by providing recommendations on top of a string analysis from the input field. Autocomplete supports the user in not just choosing existing terms from a predefined knowledge base (e.g., a thesaurus) but also helps the user to get an overview over the various contexts in which a term claims semantic validity. While the quantity of relevant concepts and labels is restricted by the scope of the domain, equivalence relations are one of the core elements within autocomplete functionalities, as they help the user to drill an arbitrary search term down to a corresponding concept. In contrast hierarchical and associative relations are of minor importance for autocomplete functionalities as information about the hierarchical depth of a thesaurus usually does not provide additional information for the construction of the search term. On the other hand information about polyhierarchies and homonyms are of major importance as they help the user to define the context in which the chosen concept demands validity.

Query Formulation / Expansion.

A thesaurus as a search tool supports query formulation and query expansion. Query terms can be widened, narrowed or translated based on the terminological pool of the thesaurus and the corresponding semantic relations. In a moderated search alternative labels (equivalence relations) and related concepts (associative relations) are used to ex-
<table>
<thead>
<tr>
<th></th>
<th>Classifying / Filtering</th>
<th>Indexing</th>
<th>Autocompletion</th>
<th>Query Formulation / Expansion</th>
<th>Recommendation</th>
<th>Glossary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concepts</td>
<td>Quantity restricted by the scope of application</td>
<td>Quantity restricted by the scope of domain</td>
<td>Quantity restricted by the scope of domain</td>
<td>Quantity restricted by the scope of domain</td>
<td>Quantity restricted by the scope of domain</td>
<td>Quantity restricted by the scope of domain</td>
</tr>
<tr>
<td>Labels</td>
<td>Quantity restricted by the scope of domain</td>
<td>Quantity restricted by the scope of domain</td>
<td>Quantity restricted by the scope of domain</td>
<td>Quantity restricted by the scope of domain</td>
<td>Quantity restricted by the scope of domain</td>
<td>Quantity restricted by the scope of domain</td>
</tr>
<tr>
<td>Equivalence Relations</td>
<td>alt/hidden relevant</td>
<td>Especially, alt and hidden relevant</td>
<td>Especially, alt and hidden relevant</td>
<td>Especially, alt and hidden relevant</td>
<td>Especially, alt and hidden relevant</td>
<td>Especially, alt relevant</td>
</tr>
<tr>
<td>Homonyms</td>
<td>Increase complexity</td>
<td>Have to be qualified</td>
<td>Have to be qualified</td>
<td>Have to be qualified</td>
<td>Have to be qualified</td>
<td>Have to be qualified</td>
</tr>
<tr>
<td>Hierarchical Relations*</td>
<td>Clear structure important</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Relevant with respect to algorithmic processes</td>
<td>Clear structure important for systematic display of thesaurus not for alphabetic display</td>
</tr>
<tr>
<td>Polylhierarchies</td>
<td>Should be avoided</td>
<td>Allowed</td>
<td>Have to be qualified</td>
<td>Not relevant</td>
<td>Allowed</td>
<td>Allowed</td>
</tr>
<tr>
<td>Hierarchical Depth</td>
<td>Depth restricted by the scope of application</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Levels needed to structure domain. Important for systematic display of thesaurus not for alphabetic display</td>
</tr>
<tr>
<td>Associative Relations</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Relevant for broadening the valid context</td>
<td>Relevant with respect to algorithmic processes</td>
<td>Relations important for systematic display of thesaurus not for alphabetic display</td>
</tr>
<tr>
<td>Definitions</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Relevant</td>
</tr>
<tr>
<td>Notes</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Relevant</td>
</tr>
</tbody>
</table>

* If the thesaurus structure provides necessary information for algorithmic processes, the importance of hierarchical and associative relations varies not just according to the application area, but also to the methodology applied to serve a specific application.

Pand the search query. While equivalence relations are well suited to define the lexical entry point into a knowledge model, associative relations help to broaden the context, in which a search query demands validity. Hierarchical relations may also be used to show alternative search terms within a given context (path dependence) but are generally of minor importance for the query construction. For better navigation, results can be sorted according to their classifi-
cation or filtered according to defined facets as a result of a previous classification (see above).

Recommendation.

A thesaurus can provide recommendations that could improve retrieval of indexed content, autocomplete suggestions or query formulation/expansion (see above) by using the domain knowledge built in the thesaurus via relations. All relation types are relevant for providing recommendations but especially associative relations and hierarchical relations play an important role because they could be used to suggest alternative search queries or help to retrieve content that is not directly related to the search terms but related to the subject of the search (e.g., using broader or sibling terms in a hierarchy) or related to the scope of the search (e.g., using related terms).

Glossary.

Glossaries can be beneficial for the user in various ways. Since the aim to completely describe the concepts of a domain all structural elements defined are relevant. A Glossary should provide a consistent and complete overview of a domain and by that could serve as a knowledge base or agreed reference of terminology for that domain. This implies also the need to clarify the meaning of concepts defined in a thesaurus by means of providing definitions, examples and scope notes. In this context, a thesaurus-based glossary can be seen as a source of metadata that can be, for example, used to provide context-sensitive help in information systems.

5. REFERENCES


