Exploring Structural Differences in Thesauri for SKOS-based Applications

Helmut Nagy
Semantic Web Company GmbH
Lerchenfelder Guertel 43
1160 Vienna, Austria
h.nagy@semantic-web.at

Tassilo Pellegrini
Semantic Web Company GmbH
Lerchenfelder Guertel 43
1160 Vienna, Austria
t.pellegrini@semantic-web.at

Christian Mader
University of Vienna, Faculty of Computer Science
Liebiggasse 4/3-4
1010 Vienna, Austria
christian.mader@univie.ac.at

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, Facetted 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/

I-SEMANTICS 2011, 7th Int. Conf. on Semantic Systems, Sept. 7-9, 2011, Graz, Austria

schematic level (See [3], [6], [5]). Its comparably low ontological (semantic) complexity makes SKOS an ideal standard to be utilized for collaborative knowledge organization purposes especially within the context of socially generated classification schemes (i.e. [8]).

With the Linked Data initiative[2] gaining momentum in the past years, SKOS (Simple Knowledge Organization System) has emerged as a common 'standard' (currently a W3C recommendation) for expressing knowledge organization systems (KOS) such as thesauri or taxonomies. SKOS features a concept-oriented approach, with a concept being "An idea or notion; a unit of thought." (as defined in the SKOS definition[3] 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[4], Eurovoc[5], GEMET[6] and STW Thesaurus for Economic[7] but also for other types of controlled vocabularies like subject headings (MeSH[8], LCSH[9] 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 applications scenarios demand different structural specificities of a thesaurus.

3. ASSUMPTIONS ON STRUCTURAL ATTRIBUTES FOR APPLICATION-SPECIFIC THESAURI

According to our hypotheses, the structure of a thesaurus influences the quality of the application output. With reference to the work of Klees & 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.

In the following we will go into more detail on the structural requirements defined for the different application scenarios.

[2]http://linkeddata.org/
[3]see http://www.w3.org/2009/08/skos-reference/skos.rdf
[4]http://aims.fao.org/website/AGROVOC-Thesaurus/sub
[5]http://eurovoc.europa.eu/
[6]http://www.eionet.europa.eu/gemet
[7]http://zbw.eu/stw/versions/latest/about
[8]http://www.nlm.nih.gov/mesh/
[9]http://id.loc.gov/authorities/

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 exploravity 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 arbritary 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 1: Structural Requirements for Different Application Scenarios

	Classifying / Filtering	**Indexing**	**Autocompletion**	**Query Formulation / Expansion**	**Recommendation**	**Glossary**
Concepts	Quantity restricted by the scope of application	Quantity restricted by the scope of domain	Quantity restricted by the scope of domain	Quantity restricted by the scope of domain	Quantity restricted by the scope of domain	Quantity restricted by the scope of domain
Labels	Quantity restricted by the scope of domain	Quantity restricted by the scope of domain	Quantity restricted by the scope of domain	Quantity restricted by the scope of domain	Quantity restricted by the scope of domain	Quantity restricted by the scope of domain
Equivalence Relations	alt/hidden relevant	Especially, alt and hidden relevant	Especially, alt and hidden relevant	Especially, alt and hidden relevant	Especially, alt and hidden	Especially, alt relevant
Homonyms	Increase complexity	Have to be qualified	Have to be qualified	Have to be qualified	Have to be qualified	Have to be qualified
Hierarchical Relations*	Clear structure important	Not relevant	Not relevant	Not relevant	Relevant with respect to algorithmic processes	Clear structure important for systematic display of thesaurus not for alphabetic display
Polyhierarchies	Should be avoided	Allowed	Have to be qualified	Not relevant	Allowed	Allowed
Hierarchical Depth	Depth restricted by the scope of application	Not relevant	Not relevant	Not relevant	Not relevant	Levels needed to structure domain. Important for systematic display of thesaurus not for alphabetic display
Associative Relations	Not relevant	Not relevant	Not relevant	Relevant for broadening the valid context	Relevant with respect to algorithmic processes	Relations important for systematic display of thesaurus not for alphabetic display
Definitions	Not relevant	Not relevant	Not relevant	Not relevant	Not relevant	Relevant
Notes	Not relevant	Not relevant	Not relevant	Not relevant	Not relevant	Relevant

* 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.

4. CONCLUSION & OUTLOOK

In this paper we outlined conceptual assumptions on the structural requirements for various thesaurus-based applications. Our analysis indicates that some application types allow to create a single thesaurus to support different scenarios (e.g., Autocomplete and Query Formulation / Expansion), while other applications demand different thesauri or a defined subset of a thesaurus to support certain functions (e.g., Filtering / Classifying and Indexing). Another result derived from the matrix is that the different application scenarios imply different complexity (e.g., Autocomplete vs. Glossary), hence differing in terms of effort and costs required for developing a vocabulary in a sufficient quality. So two main aspects have to be taken into account when developing a thesaurus:

* What application scenarios should be supported?
* What structural elements are needed to support those scenarios?

To falsify their hypotheses the authors are working on empirically testing their assumptions by taking an existing domain-specific thesaurus (e.g., STW) and simulate all defined applications scenarios with this thesaurus and a defined set of documents. For some scenarios quality measures are at hand (e.g., indexing, retrieval), while for others they will have to be developed (e.g., Autocomplete, query formulation / expansion). With this setup the usability of a thesaurus in different application scenario can be evaluated. To examine if the number of required structural elements influences the effort of the creation of a thesaurus an additional empirical study needs to be setup where several thesauri for the same domain are developed according to the different structural needs defined in the matrix for the different application scenarios. This study could be concluded by applying the quality measures for the different scenarios developed prior to the different developed thesauri.

5. REFERENCES

[1] Documentation – guidelines for the establishment and development of multilingual thesauri. Norm ISO 5964, International Organization for Standardization, Geneva, Switzerland, 1985.

[2] Information and documentation – thesauri and interoperability with other vocabularies – part 1: Thesauri for information retrieval. Norm (Draft) ISO 25964-1, International Organization for Standardization, Geneva, Switzerland, 2011.

[3] P. Avesani and M. Cova. Shared lexicon for distributed annotations on the web. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 207–214, New York, NY, USA, 2005. ACM.

[4] V. Broughton. *Essential thesaurus construction*. Facet Publishing, London, 2006.

[5] F. Echarte, J. J. Astrain, A. Córdoba, J. Villadangos, and A. Labat. Acoar: a method for the automatic classification of annotated resources. In *Proceedings of the fifth international conference on Knowledge capture*, K-CAP '09, pages 181–182, New York, NY, USA, 2009. ACM.

[6] K. Golub, J. Moon, D. Tudhope, C. Jones, B. Matthews, B. PuzoD, and M. Lykke Nielsen. Entag: enhancing social tagging for discovery. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '09, pages 163–172, New York, NY, USA, 2009. ACM.

[7] D. Kless and S. Milton. Towards quality measures for evaluating thesauri, 2010.

[8] F. Orlandi and A. Passant. Semantic search on heterogeneous wiki systems. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, WikiSym '10, pages 4:1–4:10, New York, NY, USA, 2010. ACM.

[9] D. Soergel. Indexing and retrieval performance: The logical evidence. 1994.

[10] D. Soergel. Thesauri and ontologies in digital libraries: 1. structure and use in knowledge-based assistance to users. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, JCDL '02, pages 415–415, New York, NY, USA, 2002. ACM.

[11] B. Stvilia, L. Gasser, M. B. Twidale, and L. C. Smith. A framework for information quality assessment. *J. Am. Soc. Inf. Sci. Technol.*, 58:1720–1733, October 2007.

[12] Y. Wang, N. Stash, L. Aroyo, L. Hollink, and G. Schreiber. Semantic relations for content-based recommendations. In *Proceedings of the fifth international conference on Knowledge capture*, K-CAP '09, pages 209–210, New York, NY, USA, 2009. ACM.

skos-thesauri