EXPLICIT CONCEPTUALIZATIONS FOR KNOWLEDGE MAPPING

Willem-Olaf Huijsen
Telematica Instituut, P.O. Box 589, NL-7500 AN, Enschede, The Netherlands

Samuël J. Driessen
Océ Technologies, P.O. Box 101, NL-5900 MA, Venlo, The Netherlands
Jan W.M. Jacobs
Océ Technologies, P.O. Box 101, NL-5900 MA, Venlo, The Netherlands

Keywords: information classification, conceptualizations, knowledge mapping, ontologies, semantic networks, thesauri.

Abstract: Knowledge mapping supports members of an organization in finding knowledge available within the organization, and in developing insights into corporate expertise. An essential prerequisite is an explicit conceptualization of the subject domain to enable the classification of knowledge resources. Many tools exist to create explicit conceptualizations. This paper establishes a set of requirements for conceptualization tools from the perspective of knowledge mapping. Next, a number of tools are reviewed – thesauri, ontologies, and semantic networks – using the following criteria: the complexity, the effort required, and the degree to which it is possible to integrate it into the overall knowledge mapping system.

1 INTRODUCTION

Knowledge mapping is about making the knowledge that is available within an organization transparent, and is about providing insights into its qualities. Generally, when employees look for knowledge, they draw from three sources: other employees, documents of various types, and information systems. First, other employees typically include close colleagues and other colleagues one knows to have relevant expertise. The search is typically limited to only a few people. Still, there may be others with high-quality knowledge one misses because one does not know them or does not know all of the expertise one's colleagues have. Second, documents often come in large numbers, and with a poor structure to them, making a quick and effective search very difficult. Third, information systems tend to be numerous too, and each system has a different interface and internal structure, so that finding knowledge and piecing together information from across a number of systems is a lot to ask. The distributed nature of organizational knowledge makes it very hard to get a clear, complete overview, and to draw conclusions.

Knowledge-mapping systems (KMSs) provide support for addressing these issues, collecting data on the corporate knowledge from various information systems.

The knowledge-mapping process can be said to consist of the following steps. First, raw data is acquired from one or more sources. This typically involves some *basic processing* such as filtering or keyword extraction. The resulting first-order data is stored in the *knowledge-mapping database* (KMDB). In order to obtain more meaningful information, it may be further analyzed, aggregated, and contextualized, resulting in *higher-order data*. By visualizing the first-order and higher-order data in specific ways, and taking into account user preferences, knowledge maps can be produced that provide insights into corporate knowledge.

2 CONCEPTUALIZATIONS

To talk about a domain of knowledge, one needs a way to label pieces of knowledge and the relationships between them. This is done by using conceptualizations. To establish a common frame of reference, we review some basic notions. A *concept* can be defined as any unit of thought, any idea that forms in our mind [Gertner, 1978]. Often, nouns are used to refer to concepts [Roche 2002]. *Relations*

Huijsen W., J. Driessen S. and W.M. Jacobs J. (2004).
EXPLICIT CONCEPTUALIZATIONS FOR KNOWLEDGE MAPPING.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 231-236
DOI: 10.5220/0002634202310236

form a special class of concepts [Sowa, 1984]: they describe connections between other concepts. Modifiers (or attributes) may be attached to concepts and relations to restrict or clarify their scope. One of the most important relations between concepts is the hierarchical relation (*subsumption*), in which one concept (*superconcept*) is more general than another concept (*subconcept*). *Instances* are concrete objects that may be examples of the more abstract concepts. A *conceptualization* is "the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them" [Genesereth & Nilsson, 1987]. A conceptualization can be divided into a conceptual model and an instance model. The *conceptual model* includes all concepts of interest and the relations between them. The *instance model* includes all instances of interest and the relations that link these instances to each other and to the concepts.

The conceptual model states that the individuals of interest to the organization are its employees and the external contacts. In terms of the knowledge-mapping process the instance model data is derived from the external information systems such as document management systems and organizational databases. This data is stored in the KMDB. The conceptual model typically is not available in any information system, and must be developed as part of the KMS. The KMDB invariably requires a large number of specific relations corresponding to the information presented by the various knowledge maps. Since a thesaurus can only accommodate 3–5 relations, it is not an option to integrate the KMDB and the thesaurus (Fig. 1a). However, in the case of ontologies and semantic networks, any number of specific relations can be incorporated.

3 CRITERIA

A knowledge-mapping tool uses the conceptualization in several situations: the indexing of knowledge sources, concept naming for user interaction, browsing the concept space, and relevance ranking.

Indexing determines for a given knowledge source which concepts it is concerned with. Essentially, indexing consists of examining the document and establishing its subject content; identifying the principle concepts present in the subject; and expressing these concepts in the indexing language [ISO 1985]. Therefore, indexing requires the conceptualization to include, for each concept of interest, all terms that describe it: synonyms and spelling variants. If a term is ambiguous, then hyperonyms and hyponyms, and other related concepts may be used for disambiguation.

The conceptualization is also used to name relevant concepts for communication with the user. The following must be addressed: technical terms may have any number of synonyms and spelling variants, and they may be ambiguous. Generally, this is dealt with by selecting, for every concept, from the set of synonyms and spelling variants, a single, unambiguous term as the *preferred term*, while the others are the *non-preferred terms*. This also ensures consistent naming.

An important feature of a knowledge-mapping tool is browsing the system's conceptual space. This means that the system must contain descriptions of the relations between the concepts. The number of relations depends in part on the application's requirements and on the number of instances of the relations. We find that the minimum is two: the subsumption relation and the association relation. The intended application may require certain distinctions between types of relations. A practical consideration is that if there are too few relation types, then a concept may have very many relation instances with the same name. Then it may be too difficult to find the information one is looking for.

Given a query, a knowledge-mapping tool will try and find relevant knowledge items. In presenting the knowledge items, the results are ranked according to their relevance to the query. A knowledge item may also refer to the concept in question using synonyms

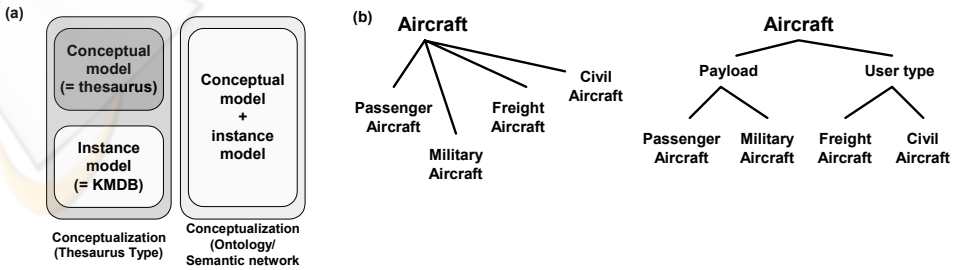

Figure 1: (a) Two conceptualization types, (b) Facets

and other related terms. Thus, relevance ranking can also benefit from knowledge about relations between terms.

Conceptualizations consist of a set of concepts on the one hand and a set of relations between those concepts on the other. From the perspective of the above-mentioned applications, we observe that having terms for the set of concepts is sufficient for indexing and concept naming. In addition, information on the relations can improve the indexing process. However, for concept browsing, both the set of concepts and the relations between them is necessary. Thus, a term set provides a minimal basis for knowledge mapping, but additional information on the relations between concepts is necessary for concept browsing.

We now summarize the requirements on a knowledge classification system. First, it must include, for every concept, all terms that describe the concept: synonyms and spelling variants, and related concepts (hyperonyms, hyponyms, etc.). Second, it must include a set of different relation types. At least subsumption and association are required. More relations are needed if the application requires so, or if the number of relation instances with the same name becomes impractical.

4 TOOLS

For knowledge mapping we consider the following three conceptualization types: thesauri, ontologies, and semantic networks. Other, simpler structures such as keyword lists and taxonomies are not taken into account as they fail to meet basic requirements. After reviewing these tools, section 5 evaluates them with respect to the criteria established above.

4.1 Thesauri

Thesauri are classification systems that combine a set of technical terms with a few basic relations. Our discussion is inspired by the ISO and ANSI/NISO standards [ISO 1986, NISO 1993, cf. Wielenga 2001]. The latter defines a thesaurus as "a controlled vocabulary arranged in a known order and structured so that equivalence, homographic, hierarchical, and associative relations among terms are displayed clearly and identified by standardized relationship indicators that are employed reciprocally".

Thesauri divide the technical terms into sets of synonyms. From each set, one *preferred term* is chosen to represent the underlying concept; the others are *non-preferred terms*. The included relation types are synonymy, subsumption, and association.

A clear advantage of thesauri is their simplicity, which allows for an easier implementation. The downside is that since only three basic relations are distinguished one may feel that many different relations get mixed up.

Of course, depending on one's application, one may vary the number and nature of the basic relations used in the thesaurus. As the hierarchical relation is the primary structuring principle, we focus on two variations that improve the clarity of the hierarchy.

An important and generally applicable variation explicitly distinguishes different types of hierarchical relations: the generic relation, the whole-part relation, and the instance relation. With only one hierarchical relation, in browsing concepts, these different types of subordinate concepts will not be readily distinguishable. Alternatively, using three relations, the presentation for concept browsing can be made much clearer.

A second variation is based on the subdivision of concepts. One can use different characteristics of a concept as the criterion for subdividing it (*facets*) (Fig. 1b).

4.2 Ontologies

Generally, an ontology is an explicit specification of a *shared* conceptualization. An ontology consists of concepts, terms, and relations between the concepts and terms [Huijsen & Driessen 2003].

An important observation about the ontology is that it defines terms in the domain and relationships between these terms in a formal way, but it does not specify the meaning of the terms. Therefore, the definition of an ontology can be focused to a formal specification of *a part of* a conceptualization. The names of the concepts and the description of the ontology in natural language are therefore important for the ability to understand and use an ontology.

In practical terms, developing an ontology includes defining concepts in the ontology, arranging the concepts in a subsumption hierarchy, defining facets and describing allowed values for these facets, and filling in the values for facets for instances [Noy & McGuinnes 2001]. In contrast to thesauri, an explicit difference is made between concepts and instances. Furthermore, while subsumption is the main relation in ontologies, there is no limit to the number of relation types that can be used in ontologies.

4.3 Semantic Networks

A semantic network is a graph of the structure of meaning. A semantic network represents knowledge as a graph. An idea, event, situation or object invariably has a composite *structure*; this is

represented in a semantic network by a corresponding structure of nodes representing conceptual units, and directed links representing relations between units. It becomes *semantic* when you assign a meaning to each node and link [Lehmann 1992]. Each concept is defined by its links to other concepts [Sowa 1984]. Concept meaning is extended by tracing outward from a particular concept to all the concepts associated with that particular concept. Semantic networks aim to represent any kind of knowledge which can be described in natural language. A semantic network system includes not only the explicitly stored net structure but also methods for automatically deriving a much larger body of *implied* knowledge. The essential idea of semantic networks is that the graph-theoretic structure of relations and abstractions can be used for inference as well as understanding [Lehmann 1992].

Semantic networks do not distinguish between concepts and instances, as ontologies do. As with ontologies, there is no limit to the number of relation types used in ontologies. The relations between the concepts/instances are bidirectional.

In an ontology the subsumption relation is the basic structure, and there typically is a strict distinction between concepts and instances. By contrast, in a semantic network, all concepts, instances and all relations have an equal status. The subsumption relation is just one of the relations, and no distinction is made between concepts and instances. This is an advantage because this makes abstraction and inference over the concepts easier. The number of relations in thesauri is fixed. In ontologies this relation is not as strict by distinguishing between concepts and instances. Semantic networks have rich relations between the concepts. This means that there are many more relations between the concepts (than in ontologies), the (semantic) distance between the concepts can be given, and finding related terms is not limited by the hierarchy (as in thesauri and ontologies).

A conceptualization in a knowledge-mapping tool is used to describe the content of knowledge sources. The results of the indexing of knowledge sources and other information are stored in the KMDB. Note that there is a commonality between the conceptualization and the KMDB: both define relations between concepts. If one allows for a large enough number of concepts and relations in the conceptualization, then these systems can be integrated. This reduces complexity and improves performance. Such a combination has a number of implications. First, since a database typically includes many relations, the conceptualization must allow for a large number of relations. Second, the conceptualization must also include the concepts of the database, such as the knowledge sources themselves.

5 EVALUATION

From the viewpoint of conceptualizations, what are the fundamental differences between the models of thesauri, ontologies, and semantic networks. Thesauri focus primarily on the textual appearance of concepts, namely technical terms. They only indirectly recognize the existence of concepts by grouping terms into synonym sets that each have one preferred term to represent the corresponding concept. Thesaurus standards limit the power to express relations between concepts in that they define only a handful of relations.

In contrast, the ontology model focuses on the notions underlying the terms: concepts and instances. The technical terms themselves are of

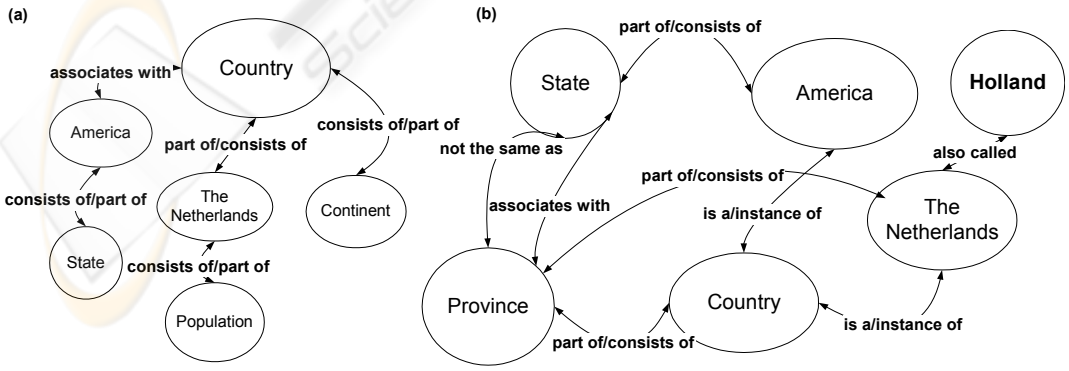

Figure 2: (a) Sample ontology (b) Sample semantic network

lesser importance. Issues like synonymy are not central, and can be dealt with in terms of the multifunctional attributes of concepts and instances. In terms of relations, the ontology model still places a focus on the hierarchical subsumption relation, but allows for arbitrarily many other relations. The ontology model thus takes a more appropriate focus and enhances expressive power by not limiting the number of relations, while still being able to deal with the issues central to the thesaurus model.

Finally, semantic networks subscribe to the more generic view that concepts and instances have so much in common that it warrants having only one type of object, also termed "concept". The distinction between the concepts and instances of the ontology model can still be made using attributes. The semantic network model also liberalizes the role of the relations. Relations are all given equal status: there is no a priori focus on the subsumption relation. Like the ontology model, the semantic network model does not limit the number of relations. In comparison to the ontology model, the semantic network model therefore is more generic: it retains the expressive power, and makes it much easier to handle concepts and instances as the same type of object where appropriate.

In summary, moving from the thesaurus model via the ontology model to the semantic network model, we see that each step corresponds to important additional insights, and each successive model is more capable to appropriately serve as a model for conceptualizations. The expressive power of each previous model is retained, not by extending the previous model, but by replacing the modeling primitives with more generic ones. As a consequence, necessary features must be softcoded in terms of the modeling primitives instead of being hardcoded as modeling primitives themselves. This means that in addition to the model itself it must be specified how to use modeling primitives such as attributes to implement the required features.

The structure of conceptualizations helps build the conceptualization. As we have seen the thesaurus is simple to build, while the ontology and semantic network call for more effort. Furthermore the structure of the conceptualization also limits or extends its inference capabilities. As for the thesaurus and the ontology, we saw that they were limited by the top-down tree structure. The semantic network is not limited by its structure.

The number of relations in a thesaurus is limited. This results in a simple conceptualization that can be built more easily. The relations in an ontology and semantic network are not limited. This implies a complex structure, especially for a semantic network. Being able to use lots of relations results in powerful inference capabilities.

Thesauri do not distinguish between concepts and instances. They use preferred terms to represent the concept. Ontologies do distinguish between concepts and instances, while semantic networks do not. The distinction between preferred terms and other terms is simple and straightforward. Distinguishing between concepts and instances, as in ontologies, is a more complex approach. In semantic networks no distinction is made between concepts and instances. This results in a complex network, but also in the advantage of being able to abstract from a concept more easily.

Higher-order processing for knowledge mapping involves some reasoning. In the thesaurus configuration, inference is hardcoded into the software and is limited by the structure of the conceptualization. Alternatively, in the ontology and semantic network configurations, we are not limited by the structure of the network. Inference typically makes use of the specific relations in the KMDB. Thus, inference capabilities of the conceptualizations only make sense if the KMDB is integrated with the conceptualization, because this enables the explicit coding of the inference rules.

Now we will judge the conceptualization based on the criteria listed in section 3. All conceptualizations can include all terms that describe the concept, e.g. synonyms and spelling variants. Thesauri do both explicitly. In ontologies and semantic networks the spelling variants are treated as synonyms. Thesauri define a preferred term, ontologies and semantic networks do not. The conceptualizations all define other relations. Thesauri usually consist of three or a limited number of relations. Ontologies and semantic networks usually also have a predefined number of relations, but theoretically can account for an unlimited number of relations.

In our view there are two viable conceptualizations for knowledge mapping: the thesaurus and the semantic network. The thesaurus variant is geared towards simplicity, ease of implementation, and reduction of work. The semantic network variant is geared towards maximum expressiveness. It has a more complex structure, which makes it harder to implement. It also allows for integration with the KMDB.

Table 1: Comparison of conceptualization Tools

	Thesaurus	**Ontology**	**Semantic Net**
Complexity	low	medium	medium
Labour intensity	medium	high	high
Integrated	–	+	+

6 CONCLUSION

Which conceptualizations should be recommended to extend a knowledge mapping system? Conclusions can be drawn from the evaluation of the conceptualizations themselves and in the light of the knowledge mapping criteria. From a technical point of view – does the conceptualization account for fully integrated conceptualization? Ontologies and semantic network satisfy these criteria, while thesauri do not. From a user perspective – how easily can a conceptualization be built? – the thesauri is the simplest and most straight-forward to build (and maintain). Ontologies and semantic networks require much more effort. However, the complexity of the semantic network is higher than that of the ontology due to the fact that no distinction is made between concepts and instances, and the possible relations are endless.

Thus, when we have to make a choice between these conceptualizations we advise to choose between the simple approach by using the thesaurus for the knowledge mapping systems or the complex approach, by choosing for the semantic network. When you want to have more complex capabilities than thesauri have, building an ontology is useless. We advise to use the semantic network approach because of the extra advantages the structure of a semantic network has and the fact that semantic networks do not differentiate between concepts and instances, which makes abstracting easier.

REFERENCES

Genesereth, M.R. and Nilsson, N.J., 1987. *Logical Foundations of Artificial Intelligence*, San Mateo, CA: Morgan Kaufmann Publishers.

Gertner, D., 1978. "On Relational Meaning: The Acquisition of Verb Meaning". Child Dev.ment, 49, pp.988-998.

Huijsen, W., and Driessen, S., 2003. "How to Build Ontologies for the Metis Pilots", Telematica Instituut, internal technical report.

ISO 5963, 1985. "Documentation – Methods for Examing Documents, Determining Their Subjects, and Selecting Indexing Terms", 1st edition, ISO.

ISO 2788, 1986. "Documentation – Guidelines for the Establishment and Development of Monolingual Thesauri", 2nd edition, ISO.

Lehmann, F., 1992. "Semantic Networks". In *Computers & Mathematics with Applications*, vol. 23, nr. 2-5.

NISO. "Guidelines for the Construction, Format, and Management of Monolingual Thesauri", NISO Press, Betheseda, MD.

Noy, N.F. and McGuinness, D.L., 2001. "Ontology Development 101: A Guide to Creating Your First Ontology", Stanford Knowledge Systems Laboratory Technical Report KSL-01-05.

Roche, C., 2002. "Form Information Society to Knowledge Society: the Ontology Issue". In: *Computing Anticipatory Systems: CASYS 2001 – Fifth Int. Conf.*, ed. D.M. Dubois.

Sowa, J.F., 1984. *Conceptual Structures: Information Processing in Mind and Machine*, Menlo Park, CA: Addison-Wesley.

Wielenga, B.J., Schreiber, A.Th., Wielemaker, J., and Sandberg, J.A.C., 2001. "From Thesaurus to Ontology",. In *First Int. Conf. on Knowledge Capture (K-CAP 2001)*, Victoria, British Columbia, Canada.