Knowledge-rich indexing of learning objects
Kabel, S.C.

Citation for published version (APA):
7 Conclusions

In the first chapter several general questions and several questions regarding each of the material handling processes were formulated. In this chapter these questions will be answered on the basis of Chapters 2 to 6.

7.1 General conclusions

In this thesis a knowledge-rich approach to re-use of material in an instructional context was investigated. First, several theoretical aspects were addressed: the material (its nature and grain size), standardization of an indexing framework, and the surplus value of ontologies for an indexing framework. This theory was followed by a proof of concept showing how ontologies are used in the material handling processes. Then several aspects concerning the use of ontologies in the indexing, retrieval and re-use processes were empirically investigated. The economic affordability of a knowledge rich approach, and the value this approach adds to the material handling processes were measured in several experiments.

7.1.1 The material

There is a large variety in electronic and non-electronic documents, as became apparent in for example the domains addressed in the IMAT project. Two important factors are relevant for the material: its grain size and nature (informational or instructional). Material with a large grain size can be source material existing in some domain, or an end-product of an instructional designer. Large pieces of material have the property that they are highly contextualized. That is, the material is developed for a specific domain, purpose, reader, and so on. Because of this context dependency, material with a large grain size is often not directly suitable for re-use and has to be segmented first. Although larger pieces of instructional material may be used by many people for some time as an end-product, such material has a low potential to be re-used in other instructional contexts than the one for which it was developed due to its contextualized nature. The notion of small re-usable chunks of material is at the heart of re-use. A general rule is: the smaller the document fragments, the larger the probability they will be re-used. Document fragments are “de-contextualized” to a degree in which they are likely to fit in other contexts. A review of the literature has learned that material that can be used or re-used in an instructional context, can be located somewhere on a continuum from informational material to instructional material. Small pieces of informational material are more likely to make creation of lesson material fast and easy, and instructional material of a relatively small grain size can be used in a flexible way by teachers and learners.

7.1.2 Standardization of an indexing framework

Several standardized annotation structures (reviewed in Chapter 2) are available for indexing and retrieval of (instructional) material. These annotation structures cover a range of different aspects of the material to be annotated and leave room for domain-dependent markup.
Standards are an important achievement. Many learning object repositories make use of standards. Different users can retrieve material from different locations based on the same, standardized annotation structure, something that was not possible 10 years ago. Current standards mainly aim at the exchange of instructional material, hence they lack the structure and expressiveness that are needed to support developers of instructional material in their tasks. Retrieval of the more elementary learning objects (information fragments) and the steps to develop lesson material from those elementary objects (such as to formulate a learning goal, strategy), are not fully supported. In this thesis a number of ontologies was investigated which can be viewed as extensions of standards.

7.1.3 Surplus value of ontologies for an indexing framework
There are four ways in which ontologies can help solving problems with defining and using an indexing framework. First, an ontology solves the problem of ambiguity and standardization of open indexes. The use of an open indexing vocabulary, for example a set of keywords derived automatically from a document, is a quick and widely used way to index (textual) material. However, keywords derived from content are not standardized. In general free use of terms in an annotation can lead to a matching problem for retrieval. Nevertheless the use of keywords is important. For example, by mapping an open vocabulary to a standardized vocabulary, keywords are used to automatically annotate material with standardized terms. Second, an ontology-based indexing approach makes it possible to annotate material from different perspectives. In this way, the scope of an annotation structure can include all aspects relevant for the material handling processes. The annotation structure described in Chapter 3 is an example of a knowledge-rich, ontology-based annotation structure. The scope of this annotation structure includes three perspectives on material: physical aspects of the material, domain-related aspects, and aspects of a task, be it an instructional design task or a task to be learned. Including these aspects in an annotation structure enables analysis and segmentation of source material into fragments, supports precision in the retrieval process, and supports the effective development of lesson material. Third, the terms in an ontology-based indexing vocabulary consist of unique concepts associated with synonyms and abbreviations. By using ontologies the problem of ambiguity of the use of natural language is solved. A fourth aspect of an ontology is structure. The terms are hierarchically organized in a structure of generic and specific concepts. A term has more meaning in a structure than in an unordered collection. The structure in ontologies is favorable for the indexing as well as the retrieval process. A structured indexing vocabulary generally results in more consistent and precise annotations. In addition a structured vocabulary allows for query generalization and specialization, which enables flexible retrieval and yields precise search results. Together these four aspects of an ontology improve the quality of the indexing process.

7.1.4 Feasibility of the concept
The theoretical framework summarized above requires a proof of concept. Is handling material using an indexing framework based on ontologies is feasible? This question is answered positively, as was shown in the IMAT project. The IMAT project provided a very realistic setting to study the entire value adding chain from source material to instructional
material. There was a real problem of teachers having to create lesson material in-house by reworking source material every time a technical manual was updated, and of students having to leaf through the technical manuals during a lesson. By providing a set of tools that break up a technical manual into smaller fragments, indexing them, storing them in a multi-media database, retrieving fragments and moving them to the preferred authoring environment, the material handling processes were significantly improved and speeded up. The applications in the IMAT project were fully based on ontologies. With the applications in the IMAT project, source material was successfully segmented, and professional educators created an instructional design, retrieved fragments and created lesson material. The developed lesson material was tried out in a real life context. The IMAT applications fulfilled a real need of users. The success of the project proved that using ontologies as a basis for the material handling processes is feasible.

7.1.5 Affordability and added value
From this proof of concept, two general questions follow: is a knowledge-rich approach economically affordable, and what is gained by such an approach?
In various experiments the costs and added value of using a knowledge-rich indexing framework were investigated. In these experiments, the use of an ontology-based indexing framework was compared with a baseline represented by the use of keywords.
One conclusion about the costs (indicated by time) of a knowledge-rich approach is not surprising: manual annotation of material using an extensive annotation structure based on ontologies is more expensive than annotating material with keywords. The costs of re-use (retrieval of fragments from a database and composing instructional material), are about the same for both methods. When using ontologies, more time is invested in generating queries, but due to precise search results, less time is necessary to judge the relevance of search results and to compose lesson material. An ontology-based approach is affordable 1) when imprecise search results are a bottleneck, and 2) when users are familiar with concepts in an ontology. If users are inexperienced, they have to become familiar with the concepts in ontologies first in order to generate queries, which requires time.
Several other conclusions can be drawn concerning the surplus value of using ontologies. In general the indexing, retrieval and re-use processes run more smoothly, with fewer frictions. Starting with the indexing process, using an ontology-based indexing framework to annotate material leads to a reduction of inconsistencies between different indexers. Retrieval is more efficient and effective. Also more relevant fragments are used in instructional products. For the re-use process, the added value of using ontologies lies foremost in sound instructional products.
In the following sections these general conclusions are refined by answering several specific questions focusing on the separate processes.

7.2 Conclusions about the analysis process
The analysis process was investigated by means of theoretical research and a case study. Two aspects were addressed: the criteria on which segmentation of source material is based, and the extent to which automation is possible.
7.2.1 Criteria for segmentation
If the grain size of material is too large to be re-used in other contexts than the one for which it was originally designed, it is analyzed and segmented into fragments first. Segmentation is usually based on three criteria: the nature of the source material, the desired grain size of the resulting fragments, and the topics treated in the source material. Segmentation of source material of an instructional nature will generally also result in fragments of a somewhat larger grain size than informational fragments, because they include instructional elements. In general, the aim is generate fragments of the smallest possible grain size, provided that fragments are coherent, and that one (main) topic is treated.

7.2.2 Automatic segmentation
Document analysis techniques highly depend on the structure in documents and the use of ontologies. Automatically discovering fragments in a bottom up way (by interpreting the layout of a document) is done using ontologies covering representational and structural properties of a document. Semantically meaningful fragments can be recognized in structured documents. Currently, automatic segmentation is possible with textual material which has a clear structure represented not only in logical elements such as chapters, sections and paragraphs but also in a topical hierarchy. Discovering fragments in a long running text would be much more difficult because the meaning of text would have to be taken into account in order to identify coherent fragments. Images as a whole included in textual documents can also be automatically detected as a fragment. Segmenting the content in images is difficult, certainly when the quality of the source material is low, but can be achieved to some degree in highly structured images such as electronic schemas. The general problem of segmenting non-textual material is still a matter of investigation.

7.3 Conclusions about indexing
Three aspects of the indexing process were investigated. The extent to which automatic indexing is possible was investigated in a case study. The degree of consistency that can be achieved in manual indexing, and the costs of manual indexing, were investigated in experiments.

7.3.1 Automatic indexing
Automatic indexing makes use of text and structure present in source material, for instance in a heading or in a caption of an image. The document analysis process in the IMAT project described in Chapter 3 showed that general and syntactic attribute values, keywords, topics, and descriptions (type and scope) can be added to a document fragment automatically. Terms occurring in documents are mapped onto concepts in an ontology and added as annotations.

7.3.2 The degree of consistency in manual indexing
If indexing must be done manually, which is currently often the case, the issue of consistency in indexing arises. In Chapter 4 it was investigated whether manual indexing can be done more consistently using an ontology than using keywords. The results showed that the
structure in ontologies has a positive effect on consistency between different indexers. Manual indexing was done more consistently using an ontology, than using flat lists of terms. Consistency between indexers was hard to achieve when an indexing vocabulary consists of value ranges such as “easy”, “medium”, “difficult”.

Even with structured value lists people did not always annotate using the same heuristics. Defining a semantic annotation structure and indexing vocabulary that is used in a standard way is problematic because people may have different mental models. The extent to which different mental models were used was partly determined by the degree of abstraction in an attribute. Abstract attributes often lack “hard” criteria to determine the attribute’s value, leading to different interpretations. The medium type can also play a role. Textual material was generally more consistently annotated than pictorial material.

Perfect consistency is hard to realize when different indexers annotate material by hand. As noted by Wason and Wiley (2000), the cataloging community ranges from the professional cataloger who describes a large number of resources on a daily basis to the casual creator of metadata who may want to provide some meaningful labels for a Web page. Together, the community of catalogers creates a corpus of metadata that contains “noise” (or inconsistencies). Using ontologies reduces this noise, as was shown in Chapter 4. Within a group of relatively inexperienced indexers with similar backgrounds, indexing without the structure in ontologies resulted in roughly about 50% noise, but using ontologies inconsistencies were reduced to about 25%. Structured indexing vocabularies not only increase consensus between indexers, but also enable retrievers of material to specialize and generalize queries. Developments in the area of indexing, for example the research conducted and indexing tools developed in the context of the ARIADNE knowledge pool system, will also contribute to making manual semantic indexing as easy, efficient and consistent as possible.

7.3.3 The costs of manual indexing

Another issue is the cost of a rich index. Chapter 6 investigated how much extra time extensive manual indexing costs with respect to indexing with keywords. The results indicate that using an extensive annotation structure to annotate material can take much longer than adding one or several keywords. Of course this depends on the extensiveness of the annotation structure. Using an annotation structure consisting of 5 semantic attributes took more than twice as long as adding several keywords. However, to think up one keyword generally took about as long as to select one concept from an ontology, which showed that indexing with ontologies is not inherently more difficult than indexing with keywords. Although there was no strong effect of medium type on the time necessary to annotate fragments, pictorial material was generally annotated faster, but the annotations showed more inconsistencies between indexers.

7.4 Conclusions about the retrieval process

The effects of using a knowledge-rich indexing framework for retrieval were investigated in several experiments. Two aspects were addressed: the added value of using ontologies for retrieval, and the costs of using ontologies for retrieval.
7.4.1 The added value of using ontologies for retrieval

The use of a knowledge-rich index improves the retrieval process in several respects. In Chapter 5 the effects of using a knowledge-rich index for retrieval and composition of lesson material were measured. Using a rich annotation structure led to more efficient and more effective search and retrieval. In addition, more relevant fragments were used in the lesson material. These positive effects were caused by a rich annotation structure (meaning that attributes representing different perspectives on material are included), as well as by the structure in the indexing vocabulary.

7.4.2 The costs of using ontologies for retrieval

The above findings confirm what is often expected when an indexing effort is made. Companies invest in indexing on the assumption that costs can be saved by reducing the time employees need to seek information. In Chapter 6, the time necessary to retrieve a set of fragments was measured. Two retrieval situations were compared: retrieval based on ontologies and retrieval based on keywords only. The results indicated that retrieval based on only keywords is somewhat faster. Typing in a keyword and hitting a “search” button was done in a very short time, whereas using several structured vocabularies to generate a query took longer. Although generating a query was done very fast in a keyword setting, it took a long time to select fragments from a list of retrieved fragments and judge their relevance. An ontology-based approach showed the opposite profile: generating a query took long and selecting fragments went fast. In an ontology setting the metadata support the relevance judgment of a retrieved set of fragments. Contrarily in a keyword setting it is necessary to inspect the fragments in order to judge their relevance.

The hypothesis that costs invested in a single indexing effort can be recovered by repeated re-use of material could not be proven in the experimental set-up, mainly because retrieval in the ontology condition took longer than was predicted. During a search action, mental effort was necessary to browse a hierarchy of terms. It can be expected that when users are familiar with ontologies, they will need less time to examine the concepts in ontologies in order to retrieve material. The point is that the benefit in terms of time of using ontologies for retrieval may only show after they have been used several times, something that was not investigated (the experiments were not repeated over time). Another way to reduce the search time is by supporting navigation, for example by providing a search functionality for concepts in an ontology. In addition, a coherent database with relatively few fragments was used in the experiments. In the case of large heterogeneous collections of material, such as the World Wide Web, retrieval based on keywords probably will take longer than was measured in the experiments. The experiments showed that using only keywords for retrieval inevitably leads to imprecise search results, even with a database with material about a single domain. Imprecise search results, in turn, inevitably lead to extra time necessary to browse search results and judge the relevance of fragments.
7.5 Conclusions about the re-use process

Two aspects of the re-use process were investigated in experiments: the quality of the instructional product, and the costs of creating an instructional product. Using a knowledge-rich indexing framework for retrieval has several positive effects on the re-use process.

7.5.1 The quality of an instructional product

In Chapter 5 it was investigated whether the use of ontologies for retrieval, when compared with keyword-based retrieval, improves the quality of the product that is created with the retrieved material. A rich index yielded a sound product; the product was complete and correct. The positive effect of using ontologies was strongest when the guideline for creating a piece of the lesson material had a clear didactic message (such as "motivate the reader"), which indicates the importance of a proper instructional design for a product of high quality. However, an ontology-based approach did not result in a better readability (in terms of coherency, redundancy and argument structure) of the product, possibly because the annotation structure used for retrieval did not support these aspects.

7.5.2 The costs of creating an instructional product

A knowledge-rich approach has the effect that costs are reduced during development of lesson material. Chapter 6 showed that, although time had to be invested in indexing and retrieval, creating lesson material was done considerably faster compared to a keyword approach. When composing lesson material from a set of retrieved fragments, about 1/3 of the time was saved with respect to a keyword approach.

7.6 Synthesis and discussion

In this thesis the feasibility and advantages of a knowledge-rich index were empirically investigated and discussed. It was shown that a knowledge-rich indexing approach is feasible, and in many respects desirable. An ontology-based index facilitates automatic document analysis and indexing, and contributes to consistency in manual indexing. The use of ontologies also improves the retrieval process; speeds up the development task, and improves an instructional product in several respects.

In the experiments described in this thesis, an ontology-based approach was compared with a keyword approach. This was done to create a baseline to compare the ontology approach with. The created keyword situation served as benchmark of a "quick and easy" indexing and retrieval method. This may have created the impression that the two approaches exclude each other, while the opposite is true. In practice, keywords should be used in combination with other structured metadata to automatically generate standardized indexes. To facilitate different preferences of retrievers both retrieval methods should be supported.

It is impossible to imagine the World Wide Web today without keyword indexing. At the same time there is a strong tendency towards the use of closed vocabularies. Metadata standards have developed in a short time over the past few years. Ontologies become more and more available on the Web. The philosophy behind the Semantic Web is to use ontolo-
gies in order to express the meaning of content. The popularity of this concept shows from ontology representation languages (RDF(S), OWL) that recently arose. The increasing use of metadata shows from numerous examples. Large companies such as Shell and Texas Instruments adopted a knowledge-rich indexing approach to gain control over their ever-increasing document flow. Also in other areas applications for the development and use of learning objects become more and more available. For instance Muzio, Heins and Mundell (2002) describe such an application and demonstrate how it allows flexible (non-sequential) composition of chunks of lesson material, and flexible use by learners (for example examining the chunks sequentially or randomly). In general these developments indicate that the more multi-media material becomes available, the more important the performance of search engines becomes, which make use of metadata. Considering the advantages of using ontologies and the fast development in the use of semantic metadata, ontologies will be the vehicle to fully realize the potential of re-use.

7.7 Further research

A number of variables have been investigated in this thesis, but in a rather specific and controlled context. More real-life studies are necessary to provide insight in the material handling processes in daily practice.

Given the expectation that the use of metadata will intensify with the increasing amount of data available on the Internet, more research on the use of metadata is required. Investigations such as conducted by Najjar, Ternier and Duval (2003) are very important to understand how people apply semantic annotation structures on a regular basis. This is necessary because many companies and institutions struggle with an indexing problem. The use of metadata for retrieval purposes should also be studied in a real life setting, to assess which attributes are actually used for retrieval. Evaluations of the use of metadata will allow improving the annotation structures where necessary, which will alleviate problems with its use.

Another research area requiring attention is the instructional design of learning objects. The LOM standard made the development of instructional material attainable for a wide public. However not everybody who has access to learning object repositories knows how to design instructional material. Wiley stressed the danger of using LOM without instructional design principles, and provided a theory for designing learning objects (Wiley, 2000). Studying the application of such a theory in practice is necessary to provide non-professional developers with more support in developing instructional material of some quality.

In addition, more research in the area of document analysis and indexing with ontologies would certainly be profitable. Automating the analysis and indexing process can save quite some time and rules out inconsistencies. Research so far has turned out to be very fruitful for companies handling large amounts of material.

Finally, to further enhance the retrieval and re-use processes, more research on the role and application of task-ontologies is necessary. In this thesis the emphasis was on the task of developing lesson material. Tasks of learners require more research, for example how learners navigate through material in learning object repositories or learning object environments. Constructivism is a philosophy of learning founded on the premise that, by reflecting on our
experiences, we construct our own understanding of the world we live in. Learning object technology has a great potential to support a constructivist approach to learning; learning object systems are structured, flexible and dynamic. Constructivist learning environments allow learners to associate instructional material with their prior knowledge and individual experiences. Ontologies of learner's tasks may be of use to support learners in "constructing meaning" when browsing through material in a repository or learning environment. Research on the role ontologies can play in supporting learner's tasks may be worthwhile.