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

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
1 Introduction

1.1 Background

The use of computers and the possibility to exchange material over the Internet have dramatically changed the world of learning and instruction over the past few decades. With the invention of computers, initially customized instructional products such as Computer Based Training applications were developed, often larger pieces of lesson material that could be used in one particular instructional context. Customized instructional products are expensive because, once created, these products are used for some time and then become outdated. The ratio of production cost to frequency of use is high. Grounded in the object-oriented paradigm of computer science, and inspired by the concept of Lego blocks, the idea to re-use small modular components of educational material emerged around 1994, called Learning Object technology (IEEE LTSC 1). When educational material is created in a re-usable form, the economics of its use change. Modular components of educational material, fitting into different instructional contexts and made accessible by means of an index, can be re-used by any number of people simultaneously. With the repeated re-use facilitated by the growing ubiquity of the Internet, cost recovery can become a reality.

E-learning can nowadays be seen as the convergence of the Web and learning on all levels, whether it is in elementary school, college, or business. Knowledge is considered a competitive advantage and a company’s most important asset. Schools, universities, governments and corporations are increasingly adopting the approach to re-use modular chunks of (educational) material. For instance, Texas Instruments, the micro-electronic firm known as the producer of electronic chips, began to re-organize the overwhelming flow of documents related to its products. In the prospect of considerable cost-savings, Texas Instruments invested in converting all their documentation to a re-usable form. Re-using portions of existing documents resulted in cost-savings of up to 70% of the cost of new documents (Tiwana, 2002). In the Netherlands, the training market research company NIDAP predicted that the use of E-learning would increase in 2003 with 18%, at the expense of traditional teaching methods such as classroom teaching and coaching. On average Dutch companies and institutions spent 12% of their budget on E-learning in 2002. In the Netherlands the ICT sector and government spent most on E-learning (NIDAP, 2002). The conditions for realizing the potential of re-using existing material for instructional purposes are investigated in this thesis.

1.2 Problems with processes underlying re-use

A large amount of electronic material is available and this will only increase in the future. For example, Google provides access to over 4 billion Web pages. The diversity of multi-media material is increasing as well: images, audio, presentations, videos, and so on, become more

---

and more available on the Web. Browsers can adequately process this diverse material for the purpose of presentation, but have no reliable way to process the semantics of the material. Search engines provide access to material through keyword-indexes and explicit links, but these do not necessarily express the meaning of the content. To support re-use, the challenge is to find a reliable way to express the meaning of the content in material in the context of a task like developing instructional material or using material in a lesson.

What are the conditions for optimal re-use of material in an instructional context? Optimal re-use requires freely and flexible accessible (learning) materials in the form of relatively small pieces of material about one (main) topic that can be fit into larger structures such as a lesson. In practice, the available material often comes in larger chunks. Thus, for optimal re-use, these larger pieces of material must be analyzed and segmented into smaller components first, based on the idea that small fragments have a higher potential for re-use than large customized pieces of material (Wiley, Recker & Gibbons, 2000). After segmentation, material must be stored somewhere. A multitude of material resides in many databases, file-systems or is otherwise stored on computers. A crucial step in re-use of material is to retrieve it. But the story does not end with retrieval; material is retrieved for a purpose. This purpose is directly related to the tasks people are performing. Effective re-use of material requires that teachers, learners, developers and designers of instructional material be supported in their tasks. In Figure 1-1 a conceptual model of the processes underlying the handling of material to enable re-use is introduced.

![Figure 1-1 Conceptual model of material handling processes.](image)

The entire process can be seen as a value-adding chain. In each process value is added to material in order to enable re-use. The promise of re-use will only come true if this entire process runs without major frictions. However, each process brings several problems and issues.

**1.2.1 Problems with analysis of source material**

In many domains source material exists which in principle can be the basis for the construction of instructional material, for example text-books, instruction videos, manuals, encyclopedia. Yet often it is not immediately suitable as instructional material. Depending on factors such as the domain, the original purpose of the source document and the expert level of the intended reader of the source document, the source document needs to be restructured and enhanced with instructional editing. In order to be able to restructure the source document, it must be segmented into fragments. This is necessary because large pieces of material cannot easily be recombined with other material. Smaller fragments are more likely to be re-used in different instructional contexts. When segmenting source material, a decision about the grain size of fragments has to be made. Ultimately this requires a difficult decision about the criteria that determine how elementary the smallest elements of a document should be in order to optimize re-usability. The nature of a source
document also needs to be considered, as this determines what type of fragments result from segmentation. Segmenting various documents can require significant effort. While the human eye can easily detect fragments in a document (chapters, sections, paragraphs), to automatically analyze and segment material, which becomes necessary when large volumes are used, is more complicated. Although it may seem that the relevance of document analysis has decreased with electronic publications (also on the Internet) because it traditionally is concerned with scanning paper documents, this is not the case. High-quality documents are exchanged in formats in which the logical structure (chapters, sections, paragraphs) is no longer present and must be discovered (for example PDF). Re-use of this material begins with the identification of fragments embedded in larger documents.

1.2.2 Problems with storage of material

Storage of material can be done in a centralized or distributed fashion. A strict central approach would hold that all instructional materials and accompanying standard metadata are stored in a central repository. This brings the advantage of control over the material that is stored and the metadata used. However, increasing the size and scope of such a system would soon become unmanageable. At the other side, completely distributed storage would mean that all materials, annotated with customized metadata, reside on many local computers. The advantage of decentralized storage is well known from the Internet. The essential property of the World Wide Web is its universality: “anything can link to anything”. But online search engines that search the whole Web would simply return too many results. Decentralized storage can bring the disadvantage of a lack of control over the (quality and versions of) material that is stored. Moreover, usually customized metadata is used, which can hamper the exchange of material between different systems and users.

In reality, the approaches to store instructional materials lie somewhere in between these two extremes. Because instructional materials are easier to find inside restricted collections, they are generally stored in learning object repositories. Some examples are the ARIADNE Knowledge Pool system, The Gateway to Educational Materials (GEM 2), and the Multimedia Educational Resource for Online Learning and Teaching (MERLOT 3). Central repositories have the limitation that material is stored at a different location than their point of origin and their point of use. People who create material lose a sense of ownership, and even for the simplest operation one must be connected to the Web. Decentralized storage is in a way inevitable but requires compromises. When material can be published by anybody at different locations, the problem arises how to assess the quality of the material. Version control is a similar problem. Take for example an instable source document such as a technical manual that is updated at least as often as the machine it describes is revised. Old fragments need to be distinguished from updated fragments, which is problematic when material is stored in a decentralized fashion. Indexing material with version-numbers and quality ratings are possible solutions. In general, when material is stored in a distributed way on many local computers, for example in separate databases, database schemas defining pos-

---

sible objects, attributes and relations must be standardized in order to exchange material. However, database schemas are often customized and additional schemas are necessary to integrate these with standardized schemas.

1.2.3 Problems with retrieval of material
"Information-glut" is the consequence of the increasing amount and diversity of material, and the need for flexible retrieval by a heterogeneous group of users increases in parallel. Simply stated, the basic problem in information retrieval is to match terms specified in a query to terms in a document or document index. This matching problem generally leads to low recall (of all relevant documents only a few are retrieved) and/or low precision (of all retrieved documents only a few are relevant). The "medium-mismatch", which means the absence of text in multi-media material from which terms could automatically be derived, makes retrieval even more problematic.

Because there are so many digital documents that can be retrieved, access to specific that material that is relevant to the purpose of re-use is crucial. In the context of re-use of material in instruction, there are two main purposes to retrieve material: to create instructional material, and to use instructional material to teach or to acquire knowledge. Each drives a different need for information that eventually leads to the consultation of a source (Byström & Hansen, 2002). In order to create or use instructional material, designers, developers, teachers and learners all must be able to retrieve relevant material easily. To optimize retrieval and to support retrievers to judge the relevance of material in the context of a task, the retrieval process needs to take the task-context into account. Task-dependent retrieval is an increasingly important topic in information retrieval research (Cool & Spink, 2002).

1.2.4 Problems with re-use of material
The fact that not everybody in the world uses the same introductory JAVA course year after year in the same way, illustrates the continuous demand for production of instructional materials. Instructional contexts in which material is used are not static but changing: background knowledge and insights of teachers and learners change, and each context may require specific materials, teaching methods and learning strategies. Moreover, when compared with traditional classroom teaching, the Internet pre-eminently facilitates "informal and incidental learning", which can take place wherever people have the need, motivation, and opportunity for learning (Marsick & Watkins, 2001). This means that, possibly as a byproduct of some other activity, knowledge is acquired driven by an instant need for information (for example in "on the job training"), increasing the demand and therefore the production of modular components of instructional material. The difficult task of designers and developers is to create components of instructional material that have enough meaning in itself to be used individually but also fit into larger instructional structures. The product should stimulate teachers and learners to repeatedly use material in an instructional context.
1.3 Key solution: index

For many of the problems sketched above, a key solution lies in an index. To provide access to digital material, it is indexed by metadata. Metadata is “data about data”, information represented in attributes like author, keywords, title, location, ISBN, and so on. An annotation structure used to index material is a collection of attributes and permitted attribute values. The possible attribute values are defined in an indexing vocabulary. Once attribute values are assigned to a piece of material, it is called metadata. In order to serve a large, distributed and heterogeneous group of users, in the past 5-7 years there have been considerable efforts to develop and standardize annotation structures for educational material (ARIADNE Educational Metadata Recommendation, 2002; IEEE standard for Learning Object Metadata, 2002).

There are several problems associated with defining and using an indexing terminology. An indexing terminology is often not standardized and not used consistently. The indexing terminology can be defined according to different viewpoints or mental models and even in different languages. A decision on one viewpoint and one language would severely limit the re-use of indexed material. Inconsistent use of terms (synonyms, abbreviations) puts even more constraints on a sound, reliable index terminology. The level of specificity of an indexing vocabulary must also be considered. With a very detailed, specific index it is problematic to find material that perfectly matches the query. Alternatively, a general index results in too much material matching the query.

A widely used approach is to index material with keywords that are derived from the document itself. As noted earlier, this is problematic for non-textual material. In addition, keywords usually regard only the subject matter (for instance what is depicted in an image). Keywords alone are not sufficient to provide task-based retrieval support because they often lack the scope and structure necessary to simulate steps in a design task. Teachers, learners, designers and developers of instructional material require additional semantic metadata that is meaningful in the context of their tasks.

To index material with semantic metadata requires human interpretation. Indexing is often done manually; a labor intensive and difficult task. Beside high costs, manual indexing may be error-prone in terms of consistency. Metadata standards for learning material ideally require indexers from different fields, for example a domain expert, an instructional designer or a teacher. A standardized indexing vocabulary does not necessarily mean that different people apply it in a consistent way. Adding semantic metadata to material by an indexer is his or her interpretation of the relation between the available attribute values and the material. From this perspective defining the standard is only half the story. The other half is to make sure that differences between indexers are minimized.

1.4 Position of this thesis

The main goal of this thesis is to investigate how the processes that are necessary to enable re-use of material, can be improved by using a knowledge-rich index. An index can be seen as the “glue” between the material handling processes, as is illustrated in Figure 1-2.
The position taken in this thesis is that a body of formally represented knowledge in an index, is the key to resolving many of the problems in each of the material handling processes separately and in its entirety. A body of formally represented knowledge is based on a conceptualization: an abstract, simplified view of the world that we wish to represent. An ontology is an explicit specification of such a conceptualization. An ontology can be viewed as a closed indexing vocabulary, consisting of hierarchically organized concepts with attributes (for example synonyms and abbreviations), relations (is-a and part-of), constraints and instantiations. A closed vocabulary allows for standardization, which is problematic with free text indexes. Used as an indexing terminology, an ontology allows for a rich expression of knowledge about a domain, material or a task.

The central hypothesis in this thesis is similar to the hypothesis underlying the Semantic Web approach. In the context of the Semantic Web highly semantic, ontology-based annotations play an important role (Berners-Lee, Hendler & Lassila, 2001). The Semantic Web, by using ontologies as indexing vocabularies, aims to add semantic terms to documents that cannot directly be derived from the content itself (Schreiber, Dubbeldam, Wielemaker & Wielinga, 2001). Software agents can reason about the information in the metadata and carry out tasks for users. The Semantic Web is not a separate Web but an extension of the current one, in which the content in Web pages is given well defined meaning and structure by means of ontology-based indexes. Ontologies are also used to enhance the functioning of the Web in yet another way: to solve the medium-mismatch, which is the problem of absence of text in multi-media material from which keyword indexes could be automatically derived (Meghini, Sebastiani & Straccia, 2001).

In this study it is hypothesized that a knowledge-rich indexing vocabulary based on ontologies improves the handling of material in several ways. The use of ontologies plays a role in each process. In the analysis process, an ontology can serve as a knowledge base for decisions that have to be made when segmenting source material. Because a structured representation of some domain in an index leaves less room for different interpretations, it is expected that a structured vocabulary, when compared with unordered, flat lists of indexing-terms, allows for consistent indexing. Problems with decentralized storage are approached by developing metadata standards and by integrating customized indexing vocabularies. A knowledge-rich indexing vocabulary also has a high potential to yield precise search results. Retrieval performance can be enhanced through providing extensive, structured task-based retrieval support. By taking the purpose of re-use into account, a rich index may even improve the quality of instructional material and cut development cost. In the end, time and money should be saved by re-using material. However, there is a downside
in the effort needed to develop and maintain a knowledge-rich indexing vocabulary, and in the effort needed to index material. It is expensive and time consuming. The economic and conceptual problems that occur in the processes underlying the handling of material lead to several questions that are addressed in this thesis. General questions regarding the material and the indexing vocabulary are:

- Which types of material can be distinguished?
- Which metadata standards are available for handling material?
- What does the use of ontologies contribute to an indexing vocabulary?
- Is handling material using an indexing vocabulary based on ontologies feasible?
- What does using a rich indexing vocabulary based on ontologies cost with respect to other methods, and at what point will these costs be recovered by repeated re-use of material?

In answering these general questions, several additional problems must be addressed in each process. Typical questions that say something about the effectiveness of segmentation are:

- On which criteria is segmentation of source material based?
- To what extent can automatic segmentation be achieved?

Manual, semantic indexing of multi-media material is time consuming and difficult. Questions addressed are:

- To what extent can automatic indexing be achieved?
- Can manual indexing be done consistently using an ontology? Do people annotate using the same heuristics, or do different mental models make consistent semantic indexing not realizable?
- How much extra time does extensive manual indexing cost with respect to other methods?

In the retrieval process, the indexes must provide access to material. Questions are:

- Does a knowledge-rich index allow for more efficient and effective retrieval than other methods?
- Does retrieval based on a knowledge-rich index save time with respect to other methods?

With respect to the effectiveness of the re-use process, the following questions are addressed:

- Does the use of ontologies in indexing improve the quality of the product that is created from retrieved material?
- Does re-use based on a knowledge-rich index save time with respect to other methods?

1.5 Approach

The approach taken to answer these questions follows three steps: a review of relevant literature, a description of a case study, and several experiments. Starting with theoretical research, Chapter 2 reviews theories, methods and metadata standards that have been developed to solve problems occurring in analysis, storage, retrieval and re-use processes. Research areas include document analysis techniques, research in the area of ontologies,
information retrieval and information seeking, and instructional design theory. Next, a "proof of concept" of a knowledge rich indexing method is given by describing a case in Chapter 3. This case covers all processes from analysis to re-use, allowing to demonstrate how different types of ontologies support the transformation of material between the processes, and support flexible, task-based retrieval of material. Chapters 4 to 6 describe experiments about using metadata. These chapters provide some insight in the effort needed to provide an index and in what users gain by flexible, task-based access. First it is investigated whether different indexers are capable of applying a standardized indexing vocabulary in a consistent way (Chapter 4). Then the added value of using an ontology-based vocabulary for retrieval and the development of instructional material is investigated in Chapter 5. The costs and benefits in terms of time of a rich index with respect to other methods are investigated in Chapter 6. Chapter 7 provides conclusions.