Efficient Information Retrieval: Tools for Knowledge Management

Katarina Stanoevska-Slabeva, Alexis Hombrecher, Siegfried Handschuh, Beat Schmid
Media and Communications Management Institute, University St. Gallen
Müller-Friedbergstr.8
CH-9000 St. Gallen
Katarina.Stanoevska@unisg.ch

Abstract

One necessary prerequisite for reusing knowledge, coded and stored in documents, are appropriate classification and retrieval procedures. Classification accompanies the process of knowledge externalisation and retrieval supports the process of knowledge internalisation in the knowledge creation circle.

In this paper we will evaluate currently available retrieval mechanism with respect to their effectiveness in knowledge management. We will then present the Q-Technology, a comprehensive classification and retrieval technology.

1 Introduction

...who controls the vocabulary, controls the knowledge
George Orwell

Knowledge has become an important resource in many organisations. The success of an organisation depends greatly on its ability to transform personal knowledge of employees and knowledge within an organisation, not explicitly belonging to an employee, into organisational knowledge. This knowledge can then be made widely available to the entire organisation and be reused when needed.

One common form of permanent storage of organisational knowledge is documents. A document is structured information intended for human perception, that can be interchanged as a unit between users and/or systems [ISO98]. Examples of documents available in organisations, which carry information as a prerequisite for knowledge creation, are handbooks for different tasks summarising process knowledge, project reports, product descriptions and others. Thus documents contain externalised, coded knowledge related to different aspects and topics of an organisation's processes and tasks.

One necessary prerequisite for reusing knowledge coded and stored in documents are appropriate classification and retrieval procedures. Classification accompanies the process of knowledge externalisation and retrieval supports the process of knowledge internalisation by enabling the capturing of appropriate coded knowledge. Thus classification and retrieval are important parts of creating organisational knowledge.

The importance of classification and efficient, as well as qualified, document retrieval has given rise to the development of different approaches and technologies for supporting this task. They differ in how far they support the transformation from personal to organisational knowledge and from organisational to personal knowledge.

In this paper we will present the Q-Technology, which provides support for automated and intelligent classification and retrieval of knowledge. This paper is divided into the following sections: Section 2 describes the role and importance of classification and retrieval in the knowledge creation cycle. Section 3 gives an overview of the most widely used classification and retrieval mechanisms. In section 4 the Q-Technology is described. Section 5 concludes with an overview of both current state of implementation and further work.

2 The Role of Information Retrieval and Classification in Knowledge Management

To know is a feature of human beings. We define knowledge as the internal state of an agent following the acquisition and processing of information [Sch98]. An agent can be a human being, storing and processing information in his mind, or an abstract machine, including devices to store and process information [Sch98].

With human knowledge we distinguish between tacit and explicit knowledge [Non91]. Tacit knowledge is person dependent. It comprises the subjective insights, intuitions, and hunches of individuals. It is knowledge, which is deeply ingrained into the context. Harris defines it as a combination of information, context, and experience [Har96]. Explicit knowledge, on the other hand, is externalised tacit knowledge, meaning tacit knowledge that has been coded on a carrier. Externalised knowledge is information. This potential knowledge is realised when information is combined with context and experience of humans to new tacit knowledge according to [Non91]. This cyclical knowledge creation process is illustrated in fig. 1.

Proc. of the 2nd Int. Conf. on Practical Aspects of Knowledge Management (PAKM98)
Basel, Switzerland, 29-30 Oct. 1998, (U. Reimer, ed.)

http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-13/

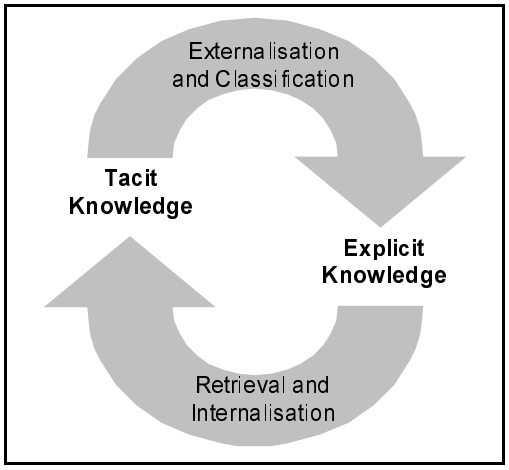

Figure 1: The Tacit-Explicit Knowledge Cycle

The creation and classification of documents is part of the externalisation process in the knowledge cycle. The top arrow in figure 1 illustrates this. After having created or acquired tacit knowledge, humans will put their ideas on paper. Today, this is often done electronically in the form of digital documents. In the next step, for the purpose of the knowledge cycle, these digital documents need to be classified so they can be retrieved at a later time by anyone interested. In order to be efficient and successful, this should be smoothly integrated into the working process. Efficient classification should be as exact as possible, yet require the least possible effort on the part of the classifier.

Retrieval is part of the internalisation process. The bottom arrow in Figure 1 represents this. When we want information about a certain topic, we ask others who we think might possess that information or we read about it. Hence, retrieval must provide the appropriate information for reuse or creation of new knowledge. In order to achieve this, only relevant answers must be supplied to a user. The retrieval of irrelevant answers, i.e. information overload, must be avoided. Therefore, answers should be as exact and as complete as possible.

In the next section we will look at different concepts for classification and retrieval systems. We will compare the systems, paying particular attention to how well the different systems aid knowledge management using the above two ideas as a criterion.

3 Overview of Existing Approaches

In this section we would like to give a brief overview of the most widely used classification and retrieval approaches. They will then be analysed from a knowledge management point-of-view.

3.1 Full Text Search

Full text search is probably the best known and most widely used search method. The general idea behind this method is to search through documents looking for key words. The best-known Internet search engines (Lycos, Infoseek, Altavista, etc.) make this very efficient by constructing an index of key words found in documents. On the Internet, so-called software robots perform this indexing, by searching through WWW pages. Constructing an indexing is nothing more then trying to capture the content of a document [Sal87]. Certain rules are applied which eliminate specific words (and, or, the, etc.) from the list of indexed words.

The full text search engines use the previously created indexes and matches them syntactically to the search words when queries are performed. As a result, the query returns those pages, which are indexed most often for the given query term or terms. Queries can be somewhat refined by using logic operators (AND, OR, NOT).

The general advantage of the full text search method is that it is very fast due to the automated indexing mechanism, which can be performed using robots. This means that classification of documents is as efficient as it can be since no human intervention is needed. Even though classification may be poor, its automation makes this part of the externalisation process, from the knowledge management point-of-view, very efficient and it integrates smoothly within the working process.

A general disadvantage of full text search is the poor quality of received results. In many cases answers returned by a full text search engine are irrelevant and incomplete, since only syntactical match is guaranteed. Not only may the results be of poor quality, meaning they are of no real interest to the user, but the number of results may be very large, which usually causes information overload. Users have to perform extensive selection activities or relevance feedback [BCr92] instead of directly consuming the information. This makes full text search a poorer approach with respect to the internalisation process of the knowledge cycle.

3.2 Case Based Reasoning (CBR)

From a logical point of view CBR is similar to metadata-based retrieval methods (see section 5) [Tho97] [Wes96]. Documents are categorised by linking them with attributes describing cases. "*A case is a contextualised piece of knowledge representing an experience. It contains the past lesson that is the content of the case and the context in which the lesson can be used*" [Watxx]. A case description usually comprises attributes describing the problem, the solution and the outcome. Cases are stored in a case base. Thus CBR is a technique of comparing the current case to a library of cases with known solutions [ISR95]. The answer contains documents, which, based on the meta-description, pertain to a specific case. This type of search mechanism is mostly used in "Help-Desk" environments [Wes96].

CBR tools, due to the categorisation mechanism, provide efficient answers. This means, unlike in the full text

search mechanism, queries will return relevant answers. If the classification is done correctly, the query will return all documents, which are relevant for a special case, not leaving any out of the result set that should logically be included. The support for the internalisation process of explicit knowledge is therefore very efficient.

Poor performance, as the amount of data becomes very large, can be a disadvantage of CBR tools. But, a certain amount of information is required within the system before useful results can be retrieved. Only when (almost) all possible types of cases have been entered in the system does the CBR tool have its greatest value. The process of updating the system with new cases is often an external process. This makes externalisation of cases (documents) very expensive and inefficient. Hence, CBR systems are the mirror image of full text search tools, with respect to the internalisation process, where externalisation is efficient and internalisation inefficient.

3.3 Metadata-based Search

The metadata-based search and retrieval method is based on meta-descriptions given for documents. Metadata is data about data. Smith [Smi96] describes it as the characterisation of information objects for the purpose of locating, evaluating, and accessing appropriate sets of objects. In database systems this is often referred to as the catalogue or schema of a database. The categorisation is achieved by adding attributes external to the content of the document, such as author or date of creation, as well as attributes describing the content of the document. The attributes have a specific semantic meaning and thus allow for, in contrast to full text retrieval methods, a semantic search. The system used in libraries is based on this principle of meta-descriptions.

Attributes should describe the properties and content of documents as fully as possible. As mentioned, since attributes can be added, aspects of the document, which go beyond its content, can be considered and used for classification. Therefore, attributes can also be used to give qualitative information about documents. For example, an attribute can describe the relevance of a document with respect to a certain topic.

Two different types of metadata-based retrieval methods are possible, depending on whether the key words are connected via clearly defined rules or whether they are freely defined. We refer to the first type as basic and the second as intelligent. The basic type is similar to the system used in libraries, where every text contains a predefined set of attributes (author, title, ISBN number, etc). By intelligent we mean that the retrieval system (the machine) can deduce information from the attributs, which enables it to guide the user in the serach process by providing alternatives for example. In this section we will focus on the basic type since it is more widely used and in section 4 we will explain the intelligent type in more detail.

These types of systems have a high descriptive quality due to the fact that the amount of attributes describing a document can be very large. Since the system can have an unlimited number of attributes for the classification, the classification can be infinitely fine allowing for a detailed and qualified search. This results in high quality answers and prevents information overload. Thus metadata-based search supports the internalisation process within the knowledge cycle more efficient than the above mentioned approaches.

A disadvantage of metadata-based retrieval systems is that prior to categorising the documents, they have to be read and understood by the person doing the classification. Thus classification is a time consuming process, which can only be performed effectively if it is smoothly integrated in the knowledge creation circle. The classification can be subjective, meaning it heavily depends on the person doing the classification. The context is different for every person and hence each person will classify things differently. No approaches have been suggested where a smooth integration of the classification process within the knowledge cycle takes place. Therefore, the cost of externalisation is very high making it inefficient.

3.4 Hypertext Systems

The idea behind hypertext systems is to link informational units in a non-linear manner [Kuh91] [Kuh92] [Kra92]. This is different from the structures found in a book. Hypertext systems form network-like structures whose connections are expressed in the form of relations. The World Wide Web is an example of a simple hypertext system.

In hypertext systems the basic search mechanisms are browsing and serendipity. Browsing is the traversing through the information in hypertext systems by clicking on links. When searching for information in this manner one tends to stumble across information of completely different value (theme). This information may then become more relevant as the one originally searched for. This is referred to as the serendipity effect. This can be either positive or negative. It is a positive effect when one is not 100% sure about the required information. The danger is that it is very easy to loose track and focus when weeding through the vast amount of information.

As explained above, when a user of a hypertext search mechanism is not certain about the information he/she is looking for, the serendipity effect of browsing can be very useful. In this manner the user can get an overview of the information available in the system. This is clearly an advantage of hypertext systems. But, overall the retrieval of information is not very efficient, as it is very difficult to find information by simply following links (without first querying a search engine). Thus, the internalisation of information within the knowledge cycle is inefficient.

The cost of setting-up hypertext systems is relative expensive. Since the relations (links) between documents can only be automated partially, most of the work has to be

done manually. Since this manual process is once again outside the working process of knowledge creation, this externalisation is inefficient with respect to knowledge management.

3.5 Overall Evaluation of Existing Approaches

At this point it makes sense to briefly compare the different technologies with each other. Once again, the evaluation criterions are efficiency of internalisation and cost of externalisation of information (knowledge) within the knowledge cycle. The two criterions are placed on two axes.

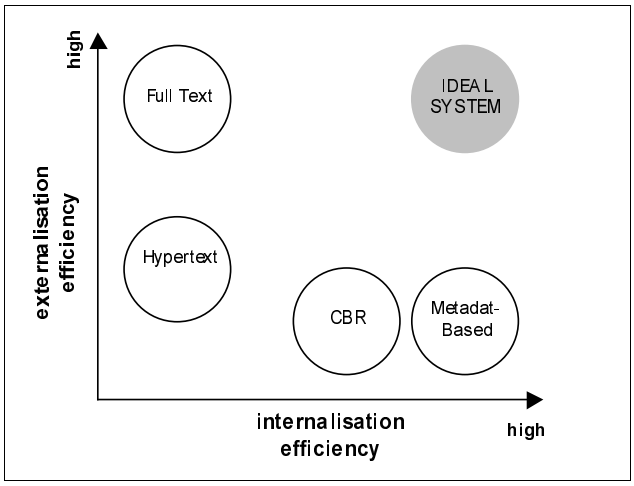

Figure 2: Evaluation of information retrieval systems with respect to knowledge management (*Note:* This figure is not drawn to scale; it just compares the relative efficiencies of the different approaches)

From Figure 2 we can see that most retrieval systems are either efficient in the externalisation or internalisation of knowledge within the knowledge creation cycle. Ideally, a system should be efficient in both processes (ideal system circle). The full text search mechanism is very efficient in the externalisation of knowledge, but very inefficient in the internalisation of knowledge. Metadata-based systems, on the other hand, are very efficient at the internalisation of information, but inefficient at externalisation. Hypertext systems, in particular the WWW, rank very poorly overall.

An ideal system for the knowledge creation cycle would be one, which is efficient at internalisation while still having low externalisation costs. In the next section we will introduce a technology, which comes close to the ideal system, the Q-Technology.

4 The Q-Technology

At the Institute for Media and Communication Management (MCM), at the University of St. Gallen, Switzerland, a generic information retrieval system has been built based on the metadata paradigm described above. Unlike most metadata-based systems, the Q-Technology allows for an intelligent information retrieval system. In this section we will explain this system and its most relevant features with respect to knowledge management.

Within the concept of the Q-Technology, the Q-Vocabulary is the means for structuring, categorising and systemising the knowledge, while the underlying Q-Calculus implements the inference mechanism.

4.1 The Q-Vocabulary

The Q-Calculus is a formal language for the description and classification of sets of objects, which was developed by Schmid [SGW96]. The basic language constructs offered are:

- *Basic sorts*, which delimit sets of objects or abstract concepts by naming them.

For example in a research paper about intelligent agents, sorts could be:

Agent, Role

- *Basic scales*, which refer to features of objects and, thus, contain the possible values of classification criteria.

For example:

Type ={Reactive, Proactive}.

ApplicationBranch = {Banking, Manufacturing, Tourism}.

- *Attributes*, which combine sorts with scales. The scale of an attribute defines partitions on the sort. In other words, it defines a classification structure for the set of objects denoted by the sort. For example: If the scale "Type" is applied to the sort "Agent" by a defined Attribute "Agent.Type", then the set of agent objects is divided in subsets of proactive and reactive agents. The denoted agents belong either to the one or the other subset.

Agent.Type = Sort: Agent ->Scale: Type

One sort can be the domain of several attributes. The Cartesian product of the attributes of one sort defines the maximal search space delimited by the sort.

Based on the above described basic language constructs more complex, i.e. derived terms can be defined:

- By using logical operators on scale elements to define and name sub-sorts of objects. Sub-sorts denote an is-a relationship to basic sorts.

For example:

Reactive Agents = Sort: Agent, Attribute: Type = Reactive

Proactive Agents = Sort: Agent, Attribute: Type =Proactive

- By applying multiplication on sorts to construct complex object sets. The components of the derived term denote a part-of relationship towards the derived term.

RoleofAgent = Agent X Role

Basic and derived terms are the foundation for further definition of new derived terms. The set of logically related terms referring to a special domain of discourse, i.e. world, forms one Q-Vocabulary. The terms of one Q-Vocabulary form a semantic net of Q-terms. The well-defined relations between the terms can be evaluated by the Q-Inference, thus allowing for complex and intelligent search. In addition each term is accompanied with a definition in natural language providing for an unambiguous semantic.

4.2 Classification of Documents with the Q-Technology

Adding vocabulary terms to a document gives it a meta-description. The vocabulary terms reflect the objects, which are considered in the document:

- *Document X {Sort* → Attribute}

The Q-Vocabulary terms form a semantic meta-layer over the documents providing a flexible and content oriented search space.

Documents are objects, which can be classified and described by the Q-Calculus as well. For example in a scientific environment documents are usually classified in *publications, abstracts, definitions*, etc. Scientific publications for example have furthermore certain classification criteria, such as:

- *Research Area := {Databases, Knowledge Management, E-Commerce}*
- *Language := {English, German, French}*
- *Audience:= {Management, Researcher, General}*

Looking more closely at the research area databases, we quickly realise that the field of databases is in itself a whole research area. Thus we can further subdivide Databases as follows:

- *Databases: {Object Oriented, Hierarchical, Relational}*

The process of describing objects in more and more detail can go on and on. The subject of relational databases is once again a complex research area. Thus, we can build complete hierarchies describing objects (c.f. 3).

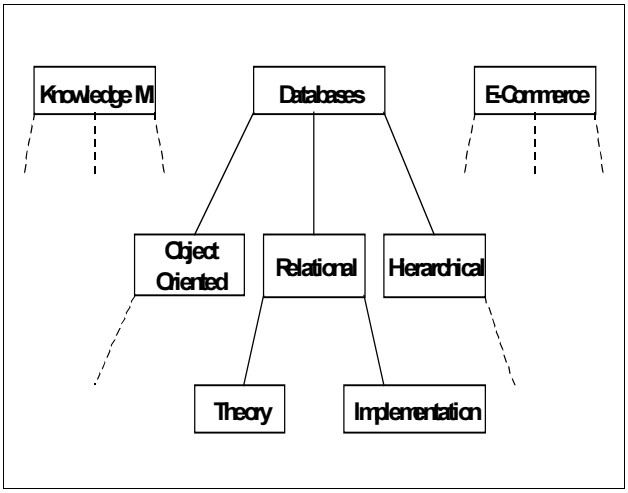

Figure 3: Hierarchical structure of documents

With the Q-Technology we are able to classify a document as precisely as we would like by choosing the most precise node in the hierarchy. By choosing one node in the is-a hierarchy automatically the whole branch up to the root is chosen. Thus one keyword provides the link to related keywords. In classical metadata-based systems, a user usually classifies a document by using keywords, which are not related. Thus keywords denoting different abstraction levels are not differentiated.

In the next section we will explain how these constructs of the Q-Technology allow for the implementation of intelligent search mechanisms.

4.3 Intelligent Search with the Q-Technology

The intelligent search functionality of The Q-Technology will be explained on the following example: A user seeks a document about knowledge management, which is furthermore written in german in 1994 (see also Figure 4).

If a search does not return a result, it can be for one or more reasons. It may be that there are no documents with all those attributes, only papers about knowledge management written in other languages then German, for example. In that case, the search request of the user returns an empty set. If one of the criterions for the failure of the search was the specification of the language, than one possibility to offer alternatives is to retrieve the documents classified with the other remaining values of the same attributes, as for example English and French, etc.

If several properties are the cause of the failure, the most irrelevant property will be determined according to user preference. We change this property while trying to keep the other properties the same. Thereby the user can specify in which order the different attributes are relevant.

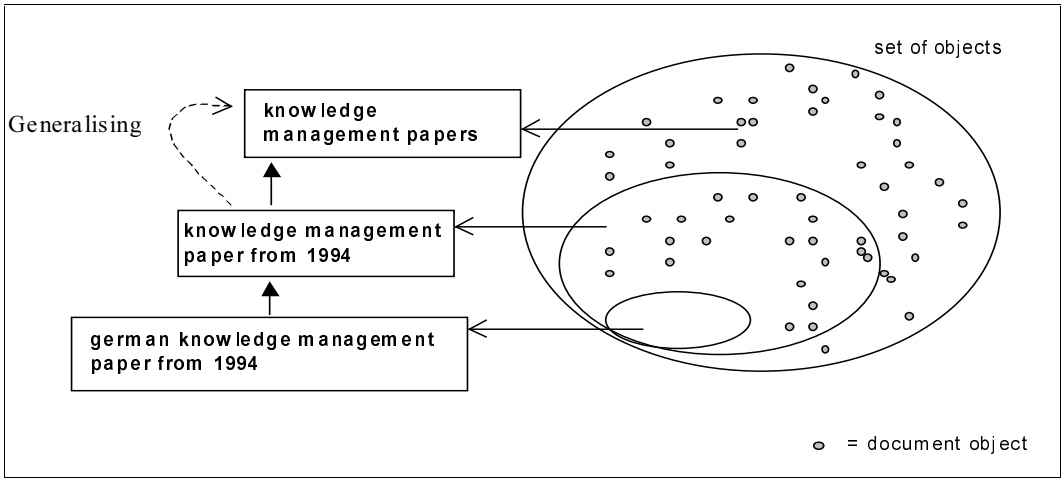

Figure 4: Intelligent search

If no object is available in the complete search area of a term, we can use the hierarchy of the terms to search in the upper nodes (c.f. 4). This method, applied to the example of the document search, would lead to a search under knowledge management documents.

5 Current State of Implementation and Further Work

The above described Q-Technology was implemented in a first prototype in Java. The prototype supports the definition of Vocabularies, the definition of their mappings to relational databases and the retrieval of documents, which are classified with the vocabulary. The query for an object originates in a WWW browser where a user specifies a query selecting one or more attributes from a list. The query is sent to the Q-Server, which resolves the attributes and matches them to a query on a database. The results are returned and the browser displays a list of documents in the form of a table. The user can then click on one of the results to view or retrieve the document.

The prototype was applied within the scientific platform available on the net the NetAcademy [HLL98] (http://www.netacademy.org). The NetAcademy stores and manages scientific publications. The Java application of the Q-Technology is used to classify and retrieve this documents.

The current Q-System only comprises one homogeneous vocabulary. Hence, repositories of documents classified according to different vocabularies can not be combined with each other. The goal for the future is to develop a system of heterogeneous vocabularies.

References

[BCr92] N. Belkin and W. Croft. Information Filtering and Information Retrieval: Two Sides of the Same Coin? Communications Of The ACM, Vol. 35, No. 12, December 1992.

[Har96] D. Harris. *Creating a Knowledge Centric Information Technology Environment*. 1996, URL: http://www.htcs.com/ckc.html.

[ISR95] Iivonen, H.; Salikovski, S.; Riitahuhta, A.: *Case-Based Reasoning and Hypermedia in Conceptual Design*. In Proceedings of the ICED'95, Praha, August 22-24, 1995, ruuvi.me.tut.fi/research/iced95_cbr.html.

[Kra92]J. Krause. *Intelligentes Information Retrieval*. in: Kuhlen, R. Experimentelles und praktisches Information Retrieval. Universitätsverlag Konstanz 1992.

[Kuh 91]R. Kuhlen. *Hypertext - Ein nicht-lineares Medium zwischen Buch und Wissensbank.* Springer Verlag 1991.

[Kuh92] R. Kuhlen. *Hypertext und Information Retrieval*, in: Kuhlen, R. Experimentelles und praktisches Information Retrieval. Universitätsverlag Konstanz, 1992.

[HLL98] Handschu, S. Lechner, U; Linke,D, Schmid, B, Scubert, P, Stanoevska, K. *The NetAcademy – A New Concept for Online Publishing and Knowledge Management*. ACOS 98 Workshop, Lisboa, Portugal 1998

[Non91] I. Nonaka. *The Knowledge Creating Company*. Harvard Business Review, November-December 1991. Pp. 96-104

[Sal87] G. Saliton. *Information Retrieval – Grundlegendes für Informationswissemschaftler*. McGrawHill, 1987

[Sch98] B. Schmid. *Wissensmedien*. Gabler Verlag, 1998 (To appear)

[SGW96] B. Schmid and G. Geyer and W. Wolff and R. Schmid and K. Stanoevska-Slabeva. Representation and automatic evaluation of empirical, especially quantitative knowledge, 1996

[Smi96] T.R. Smith. *Call for Papers: International Journal of Digital Libraries Special Issue on Metadata and Digital Libraries.* 1996, URL: http://www.llnl.gov/liv_comp/metadata/mail_dir/0027.html

[Tho97] V. Thompson. *Corporate Memories*. Byte, September 1997

[Wat98] Watson, I.: *The Case for Case-Based Reasoning* turing.une.edu.au/jirapun/watson.html.

[Wes96] S. Wess. *Intelligente Systeme für den Customer-Support – Fallbasiertesschliessen in Help-Desk und Call-Center-Anwendungen*. Wirtschafts Informatik, vol. 1 February 1996.

[ISO98] ISO 8613. *Information Processing – text and office systems – office document architecture (ODA) and interchange format*, 1998. International Organisation for Standardisat.