

Evaluation of an Ontology-based Knowledge-Management-System. A Case Study of Convera RetrievalWare 8.0¹

Oliver Bayer, Stefanie Höhfeld*, Frauke Josbächer, Nico Kimm, Ina Kradepohl, Melanie Kwiatkowski, Cornelius Puschmann, Mathias Sabbagh, Nils Werner and Ulrike Vollmer

Heinrich-Heine-University Duesseldorf, Institute of Language and Information, Department of Information Science, Universitaetsstraße 1, D-40225 Duesseldorf, Germany

Abstract. With RetrievalWare 8.0TM the American company *Convera* offers an elaborated software in the range of Information Retrieval, Information Indexing and Knowledge Management. Convera promises the possibility of handling different file formats in many different languages. Regarding comparable products one innovation is to be stressed particularly: the possibility of the preparation as well as integration of an ontology. One tool of the software package is useful in order to produce ontologies manually, to process existing ontologies and to import the very. The processing of search results is also to be mentioned. By means of categorization strategies search results can be classified dynamically and presented in personalized representations. This study presents an evaluation of the functions and components of the system. Technological aspects and modes of operation under the surface of Convera RetrievalWare will be analysed, with a focus on the creation of libraries and thesauri, and the problems posed by the integration of an existing thesaurus. Broader aspects such as usability and system ergonomics are integrated in the examination as well.

Keywords: Categorization, Convera RetrievalWare, disambiguation, dynamic classification, information indexing, information retrieval, ISO 5964, knowledge management, ontology, user interface, taxonomy, thesaurus

1. Things to know about Convera RW

At the time of writing (autumn 2005), Convera RetrievalWareTM (formerly Excalibur) has established itself as a market leader in the field of enterprise search systems. RetrievalWare is widely used in companies, government institutions, news organizations and intelligence services to “mine unstructured and semi-structured data” [2] and to afford the fast and systematic locating of relevant information. Furthermore, the system enables the efficient administration of large amounts of data in accordance with a knowledge management system and possesses the ability to handle a multitude of file formats (text, video, audio, image, etc.). Another distinctive quality that sets it apart from competing products is the option to create and imbed a thesaurus into a knowledge management framework. This form of ontology makes the implementation of a hierarchical and associatively ordered terminology possible, so that it becomes feasible to trace synonyms and resolve ambiguities.

¹This study is a result of a project of the Heinrich-Heine-University Duesseldorf. The final version of the paper was prepared by Stefanie Höhfeld, Melanie Kwiatkowski and Cornelius Puschmann.

*Corresponding author. Tel.: +49 (0)211 81 12913; Fax: +49 (0)211 81 12917; E-mail: steffihoehfeld@yahoo.de.

Libraries
 Search Results

Advanced Search Results

Viewer:

Classification Viewer

Horizontal

Vertical

Zeit

Geographie

Results (429/48)	Year 1966 (4/3)	Jahr 1967 (1/1)	Jahr 1968 (3/2)	Jahr 1969 (1/1)	Jahr 1970 (14/12)	Jahr 1971 (2/2)	Jahr 1972 (9/7)
Europa (85/19)				(1)	(3)		(1)
Amerika (92/22)	(2)	(1)	(2)		(2)	(2)	(1)
Afrika (12/9)					(1)		(1)
Naheer und mittlerer Osten und Nordafrika (12/10)	(1)						
Asien (209/39)	(1)		(1)		(7)		(5)
Internationale Organisationen (19/15)					(1)		(1)

Fig. 1. Dynamic Classification.

Convera RetrievalWare is able to catalogue large quantities of information by using so-called *libraries*. They are created – if required, topic-specifically – through the *Administration User Interface*. There are various library types available, depending on the individual application. According to the format of the documents to be managed – e.g. internet pages, database entries, text formats, data files composed with Lotus Notes, etc. – the adequate library type can be created, thus enabling access to the file data. Available library types are: File Systems for Text and Formatted Document Files, Lotus Notes, Microsoft Exchange, EMC Documentum, FileNET Panagon, RDB Libraries and FileRoom scanned documents. The *Internet Spider Libraries* offer the possibility of gathering information via a web crawler, while the *Screening Room Libraries* allow textual and visual searches within documents and images which were scanned with Convera's Screening Room application. Existing libraries are displayed in the *Search User Interface*, where a user can browse for the desired information by means of various display modes or via a classification. In turn, the classification is compiled automatically from the thesaurus, which is created and managed using the *C & C Workbench* (Cartridge and Classification Workbench); also a component of Convera RetrievalWare. At this point a special feature of Convera Retrieval Ware version 8.0 must be mentioned: the Dynamic Classification. It enables to classify search result according to the query in a dynamic way. Different contexts can be placed in connection for analysis by combining several predefined and dynamically generated classifications when presenting search results. In Fig. 1 the combination of different classification systems gets clearer, as the table, which presents the search results is ordered in a horizontal and vertical way, referring to the stored classifications.

For the rapid analysis of complete texts by means of a thesaurus, the indexed data is archived in a database. RetrievalWare currently supports Oracle and MS SQL databases. Mode, time and frequency of the indexing process can be individually configured.

Within the scope of a scientific examination, a group of students of the Heinrich-Heine-University Duesseldorf (Germany) conducted an evaluation of Convera RetrievalWare 8.0 over 10 month (October

2004 – July 2005). Because of the complexity of the system, the group restricted the analysis to the library type “File Systems” and the processing of text files and formatted document files. The main focus was on the construction (incl. implementation) of a sample thesaurus as well as on the procedure for importing the “Standard Thesaurus Wirtschaft” (abbr.: STW; [German] Standard Thesaurus of Economics). Furthermore, tests concerning data management, support of different file formats and the usability at all levels of the system were conducted. The individual settings and modifications of the system were customized in accordance with the Convera RetrievalWare documentation. This publication intends to give a comprehensive representation of our findings in the course of the project. An introduction to the components and functions of the software is provided along with a critical review of the system’s strengths and weaknesses.

In the following sections we will describe the *Search User Interface* and its available settings. We will analyse the technology and modes of operation under the surface of Convera RetrievalWare, with a focus on the creation of libraries and test thesauri, and the problems posed by the integration of an existing thesaurus. At the end of this study, we will take a look at broader aspects such as usability and system ergonomics and present the results of a poll taken in our course, pertaining to the ease-of-use of the software.

2. Search user interface

A first look at RetrievalWare’s *Search User Interface* yields few surprises. The interface is browser-based and quite similar to that of any major search engine. A single form field can be used to enter a search query. Apart from this *Basic Search* function, the *Advanced Search* option offers additional possibilities for searching (see Fig. 2). Whereas the former one’s functionality can be compared to most of the popular internet search engines, the latter is divided into three different subtypes which are covered later.

2.1. Hit list

After performing a search query, the user receives a hit list containing a title, an optional summary, a relevance bar and a file-type icon for each retrieved item. Having chosen one of the items, the user can decide on several types of data representation, namely “none”, “folders” and “table”. Since these descriptions are hardly self-explanatory, a short comment describing them would be quite helpful. After selecting, for example, the “table” view, one is confronted with two numeric values listed at the top of the table, again without further explanation. The first indicates the total number of document links available in the table. The second represents the actual number of documents classified in the table.

The relevance bar and the icon in the hit list lack a context menu explaining what precisely is presented (this is particularly true regarding the icon used for HTML files). Furthermore, the relevance bar can easily be misinterpreted, because there is both a specific relevance to the search query as well as a relative relevance assigned to similar documents. In this context we encountered the problem that even documents containing the exact search term(s) in the correct order are not rated as 100% relevant. Another problem, in our opinion, is the fact that results are not presented in the common form of quasi-concordances, i.e. with the search terms in their sentence context. A representation similar to the one used by Google [3] or AltaVista [1] would be beneficial. It is possible to receive a summary of each of the hits, but this seems to be just an excerpt from the first lines of the document without any connection to the search term. Again we regret the lack of a contextual presentation of the results.

Advanced Search

▶ Additional Query Fields

▼ Primary Search Options

Query Type

Concept

Pattern

Boolean

Query Language

English

Concept Expansion Level

Simple Variations

Exact Matches

Simple Variations

Most Strongly Related Concepts

Strongly Related Concepts

Weakly Related Concepts

select all clear all

English

German

Maximum Documents to Return

100

Match Query Language

▼ Query Term Expansions

Query Term

INTERNATIONAL refresh expansions

select all clear all

Select Expansion Terms

ADJECTIVE: international (English: SUPRANATIONAL, INTERNATIONALISTIC, INTERNATIONALIST, GLOBAL, WORLDWIDE, PLANETARY, INTERNATIONALITY, INTERNATIONALISM, NATIONAL, FOREIGN, INTERNATIONAL)

ADJECTIVE: external; outside; international (English: FOREIGN, OUTER.

Selected Terms

INTERNATIONAL; ADJECTIVE: international; SUPRANATIONAL, INTERNATIONALISTIC, INTERNATIONALIST, GLOBAL, WORLDWIDE, PLANETARY, INTERNATIONALITY, INTERNATIONALISM, NATIONAL; FOREIGN, INTERNATIONAL)

Fig. 2. Advanced Search.

2.2. Search types

Because *Advanced Search* offers a number of relevant features, we will present it here without further commenting on *Basic Search*. The latter should pose no real difficulty to most of the users, apart from the previously discussed issues. As we have already mentioned, *Basic Search* is largely similar to common internet search engines, both in terms of interface and functionality. Besides the widely used *Boolean Search* option, the system offers two interesting additional methods within *Advanced Search*: *Concept* and *Pattern Search*. In the *Advanced Mode* there are several adjustment possibilities, some of which are not easy to understand. Again a small context menu would be very useful, e.g. for the available search types. Although one is able to realize the differences between these options simply by trying them out, the novice user would appreciate additional information.

The *Concept Search* method can be used to get an automatic grammatical or semantic expansion of a term. This approach can, for example, find strong inflective forms like “stood” if the query is “stand”, and synonyms or related terms (“run” – “walk”). *Pattern Search* tolerates spelling errors and variations: a user searching for “economy” will automatically receive documents containing “economy” (and vice versa). What we miss within *Concept Search* is a quick note explaining what kind of wildcards are available and how precisely this method works. It would also be helpful to know whether or not spelling errors are individually indicated, as it is accomplished by *Pattern Search*. One drawback of *Pattern Search* is the lack of wildcards support (e.g. truncation methods). Throughout the entire system (but inside the documentation) there is no definite list of wildcards and their respective functions. A simple reference to the appropriate page in the help desk section would be useful. Within *Boolean Search*, an enumeration of the possible operators would be of interest (beside the popular ones like “AND”, “OR” and “NOT”), because many systems differ in the number and type of valid operators (for example, which is the appropriate operator for adjacency?). Perhaps a radio button to indicate the desired operator would be easier to use. It should also be pointed out that it is possible to mix the query types at least to a certain extent. By using an exclamation mark (“!”) within the search field, one can retrieve a semantic expansion of a term when using *Pattern* or *Boolean Mode*. If, for example, *Concept Search* is chosen, the tilde character (“~”) triggers a pattern expansion. While these options are available, they are not visible to the first-time user – one has to experiment or consult the documentation. In addition to these three search methods (*Basic*, *Pattern* and *Concept Search*), an information professional would appreciate further means of retrieving data, for example in the form of a simple query language en par with what is available in common database applications.

2.3. Critical consideration

Obviously the aspects mentioned above do not cause enormous problems, but some of these simple obstacles can frustrate unpractised users so much that they are prone to ignore the vast possibilities of *Advanced Search* and rely solely on *Basic Search*.

3. Technical aspects

One of Convera’s most distinct features is, as mentioned, its ability to match search patterns with the contents of source documents on a conceptual level. Instead of simply comparing surface forms, Convera RW attempts to identify the abstract concept that is referenced by a word or phrase and looks for the concept in the thesauri available. In Convera’s jargon, an *expression* is a word or phrase that refers to a conceptual entity, while a *term* is a word or phrase (i.e. an *expression*) in a particular language. Terms belong to grammatical categories such as nouns, verbs, adjectives, adverbs, prepositions, abbreviations, etc. In contrast to this, a *synset* is a language-independent concept. It may have links to *terms*, other *synsets* and *taxnodes*. If *synsets* are linked to other *synsets* they form a mesh of interrelated concepts called a *semantic network*. Wordnet [8] is an example for such a *semantic network*. A *taxnode* is also a language-independent cluster of concepts. It may have links to *synsets* and other *taxnodes*. *Taxnodes* are linked to one another in a hierarchical fashion to form a *taxonomy*. A *cartridge* is a database comparable to a thesaurus, consisting of *terms*, *synsets* and *taxnodes* and the links between them. These links have weights to indicate their relevance which are defined inside the taxonomy utilized (for more on weights, see Chapter 4 *Taxonomy and Classification*). When several *cartridges* are combined, the result is called a *super-cartridge*.

3.1. Latching

Latching is the process of expression-to-term lookup. *Expressions* in a document are associated with matching *terms* in the *supercartridge*. The process of *latching* begins by identifying *potential latches*. Any *supercartridge term* the normalized expression in the document is exactly identical to is considered a potential latch. For instance, if the document contains the word “beer”, the supercartridge terms “beer”, “BEER”, “Beer” and “bEer” would all be considered potential latches. The term “bear” would not be considered a potential latch. A potential latch has to meet the following requirements to become a true latch:

- The term’s language has to match the specified language.
- The terms part-of-speech has to match the specified part-of-speech.
- The term has one or more links to *synsets*.
- The term’s *cartridge* must be flagged as ACTIVE.

Once a latch has been confirmed, its occurrence, i.e. its location inside the analyzed document, is saved. The end result of the latching process is a so-called *synset vector*. It contains a list of *cartridge*, *synset*, *taxnode* and *occurrence* records.

3.2. Categorization

Latching is part of the *categorization* process. During *categorization*, *synset vectors* are extracted from a document to serve as a sort of summary for later use. Because *synset vectors* relate to abstract concepts, they are language-independent. What is still necessary, however, is a collection of *cartridges* that the document can be tested against to provide the vectors and allow a successful *categorization*.

3.3. Normalization

The first step in *categorization* is called *normalization*. The goal of this process is to transform the *expressions* in the *supercartridge* and the source documents into a comparable form. The numerous logically independent operations performed toward this end are referred to as *normalization*. Among these operations there is, for example, the reduction of plural forms to singular forms, i.e. the reduction of an *expression* to its word stem, a process known as stemming. Stemming also includes the removal of suffixes like -s, -ed, -ing, etc. While this method is highly applicable for English, a language such as German causes greater problems. In addition to this, there are character mapping rules which are able to deal with diacritical marks like the German Umlaut: for example “ä” may be mapped to “ae”. All letters are converted to uppercase. Punctuation characters are converted to spaces. Multiple sequential spaces are reduced to a single space. Finally, leading and trailing spaces are removed.

3.4. Idiom detection

In order to speed up research, a data structure is used which is called a *hash table*. It contains pointers to each term’s normalized expression which functions as keywords. Thus “CAR” would be used as a key for “Car”, “car”, “caR”, “cars”, etc. This method helps to save time. However, in this context specific construction such as *idioms* can create further problems.

An *idiom* is a phrase, construction or expression that is recognized as a unit in the usage of a given language and either differs from the usual syntactic patterns or has a meaning that differs from the literal

meaning of its parts taken together. An example is: “It is raining cats and dogs”. The time required for a linear search is proportional to the number of terms, which is very slow for all but the smallest *super cartridges*. Therefore a word-by-word search would take too much time and lead to wrong results. It must be clear a priori how to group words for expression-to-term lookup. Suppose, for example, one expression in the cartridge is “white polar bear” and you encounter something like “the magnificent white polar bear is now nearly extinct” in a document. If you look up one word at a time you will not latch because the hash code for “white polar bear” is not the same as for “white”, “polar”, or “bear”. It is possible, of course, to look up every existing combination of words up to the length of the longest term, but it would likely be slow. In such cases the system needs to know which words belong together.

As cartridges are loaded, the first word of each expression is hashed (put into a hash table) and all expressions with the same first word are linked together, sorted in descending order of their length in words. Document words are placed one at a time into an expression queue. When the length (in words) of the queue equals the longest expression in the *super cartridge*, one performs a hash lookup using the first word in the queue. If there is a match one then traverses the linked list of expressions (each of which has that same first word) and compares each one to the expression in the queue. Since the expressions are in sorted order, the longest match (in words) is guaranteed to be detected first, and one stops looking once a shorter expression is encountered. This prevents latching “white” as soon as something like “white polar bear” is found. Only one expression can match, though the expression can be linked to multiple terms having different languages, parts-of-speech, or unnormalized forms. Once the expression comparisons are complete the first word is removed from the queue.

3.5. Disambiguation

Another characteristic of language that causes some trouble is that of ambiguity. A term may have various meanings. This phenomenon is called homonymy. For example if we have the term “Washington”, we cannot know if it denotes the first president of the U.S.A., the nation’s capital, or the western state, without taking into consideration the context in a document. Technically this would be a term that is linked to more than one *synset* (concept group). A term that belongs to one *synset* only, without links to other *synsets* or *taxnodes* can be regarded as being of minor importance. Only if it can be seen as a part of a semantic network it is considered relevant. In the aforementioned example, “Washington” would rather be associated with the city if there are links to other *synsets* containing expressions such as “DC”, “District of Columbia” and “Beltway”. Disambiguation occurs first on the *synset*-, then on the *taxnode-level*. *Taxnode-level* disambiguation uses the depth of the taxonomy to determine which possible context of meaning is more likely to be the intended one. If a taxonomy has a multitude of hierarchical levels, the disambiguation process is likely to succeed unless there are not enough terms surrounding the ambiguous one to indicate the intended meaning (i.e. a document containing “Washington” only, would not allow successful disambiguation).

3.6. Critical consideration

Convera RW’s capability to conduct searches by analyzing documents using domain-specific taxonomies is a very promising feature. However, there are several critical points which make the implementation of these taxonomies a complicated task, partly because of design-specific problems, and partly because the task of interpreting human language accurately is extremely difficult.

We have identified the following problems:

- Convera RW is only as good as the cartridges it uses
- Language ambiguity is difficult to overcome
- Not all abstract concepts are effectively language-independent
- Variety in character-encoding is a persistent problem
- Selecting documents which are ‘tailored’ to the cartridges available is necessary for good results

3.6.1. *Convera RW is only as good as the cartridges it uses*

Convera RW can only accurately categorize and rank documents when it is using a comprehensive taxonomy in which terms are organized into many hierarchical levels. There are several reasons why a significant depth inside a cartridge is desirable:

- (a) it is much easier to browse the tree structure in RetrievalWare if there are more than just two or three levels,
- (b) related “lower” and “higher” terms factor into the ranking process significantly – a document will not be ranked correctly if no related matching terms can be found on other taxnode levels.

Finally, some cartridges are definitely easier to assemble than others. The *Time* and *Geography* cartridges pre-packaged with our Convera RW edition are much simpler in their structure than a cartridge for economics or art would have to be.

3.6.2. *Language ambiguity is difficult to overcome*

As with the categorization process as such, taxnode-level disambiguation is only effective when using a large taxonomy with many hierarchical levels. Ambiguities can be resolved if other terms in the vicinity of the ambiguous one can be found on a related level of the taxonomy. If the taxonomy lacks depth, disambiguation is likely to fail.

3.6.3. *Not all abstract concepts are effectively language-independent*

It is important to note that Convera RW is only language-independent in the sense that one can use any single language per implementation, but not that one can correctly categorize documents in a number of different languages with one cartridge. It is somewhat misleading to assume that the concept and the expression of something in a given language exist in complete independence from one another and that one can easily match a universal concept with its concrete realization in any language.

3.6.4. *Variety in character-encoding is a persistent problem*

Because of the lack of standardization in character-encoding, Convera RW is bound to make mistakes when faced with inconsistently encoded documents. Universal use of Unicode could do away with this hurdle sooner or later, but until that happens, it will continue to be an issue.

3.6.5. *Selecting documents which are “tailored” to the cartridges available is necessary for good results*

While it is certainly possible to run documents in a wide variety of formats through Convera RW, it is most unlikely that a completely accurate categorization is possible if the data used is unsuitable. If, for example, a newspaper article about the “moral debt of the United Nations in the wake of the Rwandan genocide” was parsed using a cartridge for economics, the result would probably be a match for the term “debt” with its common meaning. This does not change the fact that Convera RW can be an extremely powerful tool for knowledge management, but it highlights that language ambiguity is a problem that remains unsolved.

4. Thesaurus and classification

One special advantage of Convera RW is the ability to create and embed a thesaurus. In the course of this project we focused on testing the functions for creating new thesauri and importing existing ones. In order to test the software's import capabilities we used the German "Standard Thesaurus Wirtschaft" (STW). The *C & C Workbench* (see Fig. 3) is the tool of the RetrievalWare, which enables users to produce monolingual and multilingual thesauri (in Convera RW's jargon: *Cartridges*) and taxonomies for the support of the enquiry with Convera RW. The *C & C Workbench* consists of several tools: the *Cartridge Editor*, the *Classification Editor* and additional applications for the creation and modification of cartridges. The *Cartridge Editor* serves for the production of the cartridges which are used as thesauri. It is possible to use an existing thesaurus compliant with the ISO 5964 standard [4], as we have done with the STW.

4.1. The procedure of thesaurus creation

The *Cartridge Editor* is the tool required to create a new Convera cartridge. Internally, cartridges serve as taxonomies which contain hierarchical categories according to which the source data is classified. To each of the categories several data sets can be attached which represent the descriptors and non-descriptors of a thesaurus. The *Cartridge Editor* makes no distinction between descriptors and non-descriptors, but terms and synonyms can be weighted differently within a data set. With the help of the *Canvas* (see Fig. 3) – a work surface which vaguely resembles Microsoft PowerPoint – relations between individual data sets can be defined. By default different weights are assigned to the three relations *RT* (*related term*), *BT* (*broader term*) and *NT* (*narrower term*). The default values, which indicate the relevance for the different types of relations are: $NT = 80\%$, $BT = 50\%$ and $RT = 40\%$. Synonyms get a predefined weight of 100%. Additional types of relations with various relevancies can also be defined. Lastly, global settings related to metadata, support for different languages, weights of synonyms and applications for the whole cartridge can be adjusted.

4.2. The *C & C Workbench* put to the test

Initially, a list of the terms is either created directly inside the *Cartridge Editor* or imported manually. The *Cartridge Editor* categories have to be created separately one after the other and the terms and synonyms need to be registered and linked with relations in each data set. Using the *canvas*, term relations are mapped visually with arrows which connect a single item to other items on either side. Once all the necessary modifications have been made, the cartridge can be evaluated statistically ("Taxonomy QA Metrics") as well as exported. In the case of our test cartridge, two data files had to be created. The first one was the classification needed to later represent the search results within Convera RW. Via a runtime cartridge the *Classification Editor* was provided with the taxonomy that had been previously assembled in the *Cartridge Editor*. After another round of adjustments to the classification we exported our first data file, enabling us to produce the actual thesaurus. Convera Retrieval Ware needs both files in order to function and the sequence of the steps should not be changed. It is also recommended to use the pre-defined directory paths, as the system does not react favourably to custom settings in this area. The benchmarking tools (*Benchmark Taxonomy* and *Benchmark Expansion*) of the workbench can be used optionally for later fine-tuning.

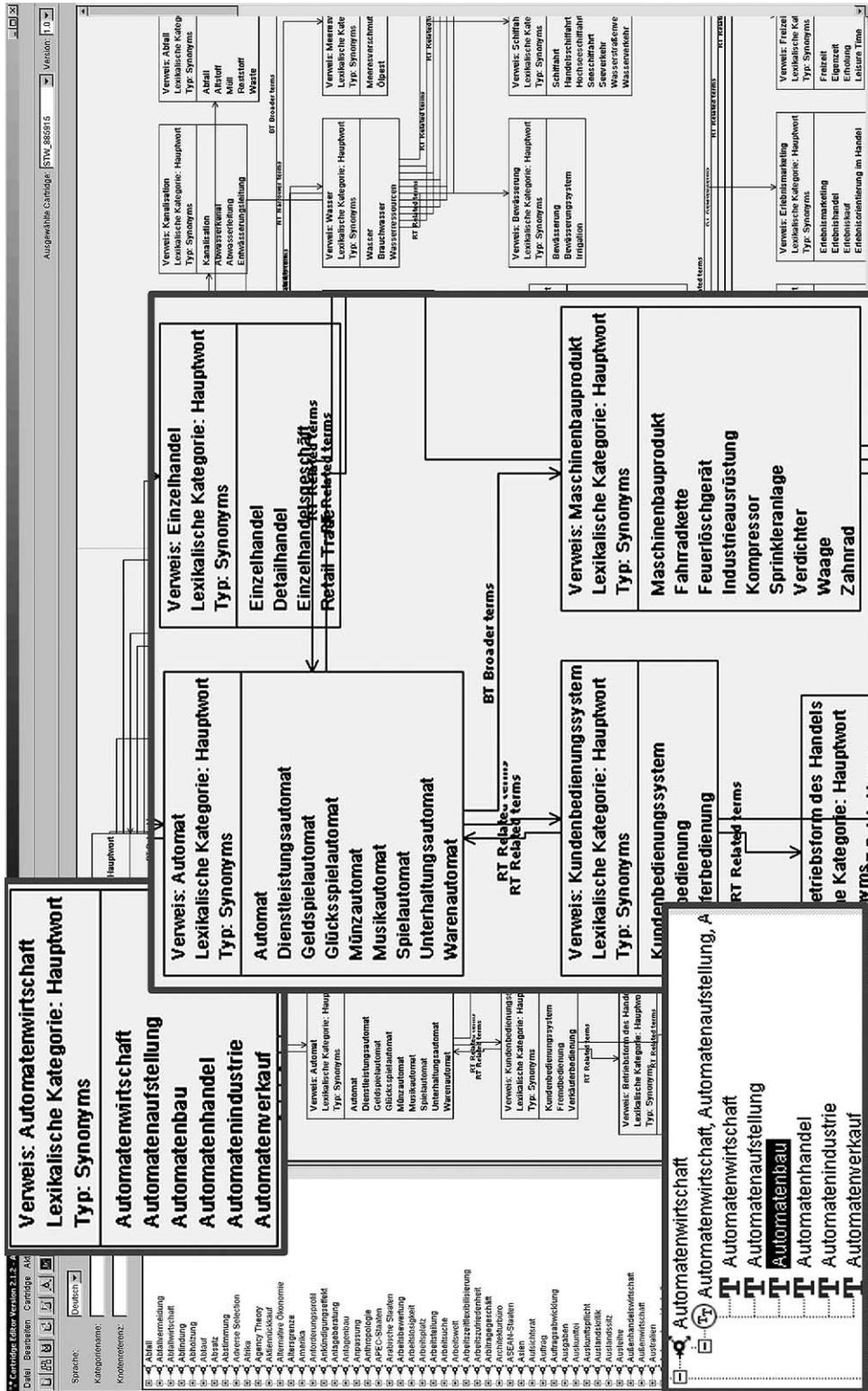


Fig. 3. Cartridge Editor of the C & C Workbench.

After creating a new thesaurus from scratch, we turned to the task of importing an already existing, complete one – the STW. To import a thesaurus into RetrievalWare’s database it should be available in the ISO 5964 format and we would like to take the opportunity to elaborate briefly on the subject.

The international standards definition ISO 5964, which contains guidelines for establishing and developing multilingual thesauri,

“should be used in conjunction with ISO 2788 and should be considered as an extension of the scope of the monolingual guidelines. Generally, the majority of processes and references embodied in ISO 2788 do possess the same validity for the multilingual thesaurus” [4, p. 1].

According to ISO 5964 these markers have to be used when defining relations: *BT* (*broader term*), *NT* (*narrower term*), *RT* (*related term*), *USE* (*use*), *SN* (*scope note*) [4, p. 4].

One problem that we encountered in this process was the fact that the STW datasets contained variables and characters which were not consistent with Convera RetrievalWare. Our method of resolution was to search, replace and erase non-consistent variables, which we did by means of a simple macro-script that was able to find and substitute certain characters. Special data-rows which – though useful – were not compliant with the ISO 5964 format caused significant problems during the import process. The consequence of such a faulty parsing is a corruption of the datasets and their attached relation markers – something that has to be avoided at all cost for obvious reasons.

After bringing the thesaurus into an ISO compliant form we imported the STW into the Cartridge Editor. When converting data from the taxonomy to a classification we experienced the following problems:

- The classification was not benchmarked completely.
- The STW has more than 1000 descriptors on the top level.
- We were surprised to find that the items of our classification were not shown in alphabetical order. Furthermore, all terms with the initials “R” and partly those with “S” were missing.

To solve the latter problem, we benchmarked the taxonomy with another version. The newly created classification was now structured alphabetically but all terms following the initial “T” were still missing.

Working with more than 1000 descriptors on the taxonomy’s top level in the Classification Editor proved to be very difficult. With the Classification Editor it is possible to build a classification automatically and then draw terms from the taxonomy into the model by hand. Since too many descriptors on the top level make quick browsing within the classification virtually impossible for the end user, we decided to integrate the STW’s seven pre-defined subject areas manually. The procedure’s advantage is that terms from the taxonomy can be assigned to multiple categories in the classification as prefigured by the thesaurus. This implicates additional efforts, as every single term has to be “tackled” and drawn into the corresponding category. On the retrieval side, it means that it is possible to browse specific subject areas and categories (as they are found in a directory) if the user is familiar enough with the thesaurus to know where to search for an individual item.

4.3. Critical consideration

Having thoroughly examined the *C&C Workbench*, we would like to discuss several points which could be improved. To guarantee a smooth workflow the client computer should meet the system requirements, otherwise operations are slowed down considerably. In the course of our project it became clear that the Cartridge Editor is not suitable for the production of larger thesauri because all changes to an existing model have to be conducted manually. It is likely that the considerable size of the cartridges

that we used was responsible for some of the error messages which we encountered. In addition to these issues, we made the following observations:

- Only classifications can be combined, cartridges (and therefore also thesauri) cannot.
- The thesaurus as a whole must either be produced manually or imported from an ISO compliant source.
- Support for multiple languages is not sufficiently documented. English is automatically preferred when German/English thesauri are created, even if only German terms exist. This makes it difficult to distinguish data sets in multiple languages when looking at the canvas.
- When creating a new taxonomy, the data sets of the categories are not assigned to their corresponding relations automatically – all relations must be drawn by hand.

The user interface of the Cartridge Editor lacks important features. There is e.g. no warning message when duplicate entries are created within a taxonomy; such duplicates can only be located at a later point using the benchmarking tool. Larger taxonomies can not be browsed in their entirety and the Editor offers too few possibilities for structuring them. Editing extensive taxonomies is tedious because of the need to frequently click and scroll and the whole process of recording data is too cumbersome. Hotkeys and a more customizable interface would simplify the process significantly. Including a feature such as “remember last input” for repeated similar actions would be a simple but effective improvement. Other simple additions such as, for example, the ability to select multiple items using the SHIFT and CTRL keys would speed up the work considerably.

Summing up our findings, we conclude that producing large thesauri and taxonomies with the editor is laborious. One may fall back on the possibility of importing a file with tabs for indentation or an external recorded thesaurus according to the ISO 5964 standard. Theoretically it is also possible to write the data sets directly into the MS-SQL or Oracle database, but since the database structure is very complex this would only make sense for very large projects. After having put the data into the database, one could employ the other tools of the editor to do the “fine tuning” of the recorded data.

5. Library configuration

Once a thesaurus is successfully created, it can be used together with *libraries*, which contain the documents that are indexed and searched with Convera RW. The number of libraries may depend on the different types of documents that they contain. The standard type of library is “File Systems” for text and formatted document files. It can handle the various Microsoft Office formats like PowerPoint, Word and Excel, but also PDF documents and a broad range of other common formats. Convera RW offers several other library types, as we already mentioned at the beginning. In the course of the project only the standard “File Systems” were used.

A library can only contain one type of data source. For example, if someone wants to use Convera RW for File Systems and the import of Lotus Notes data, one would have to create at least two libraries, one for each type of data source. Each library can have several different document sources, which have different paths for indexing. For each document source a particular default language can be chosen, so that a single library can include documents in different languages from several document sources. While setting up the library through the *Administration User Interface*, one should already know if a thesaurus will be used later on and select the appropriate field in the menu, as the settings for categorization and classification cannot be modified afterwards.

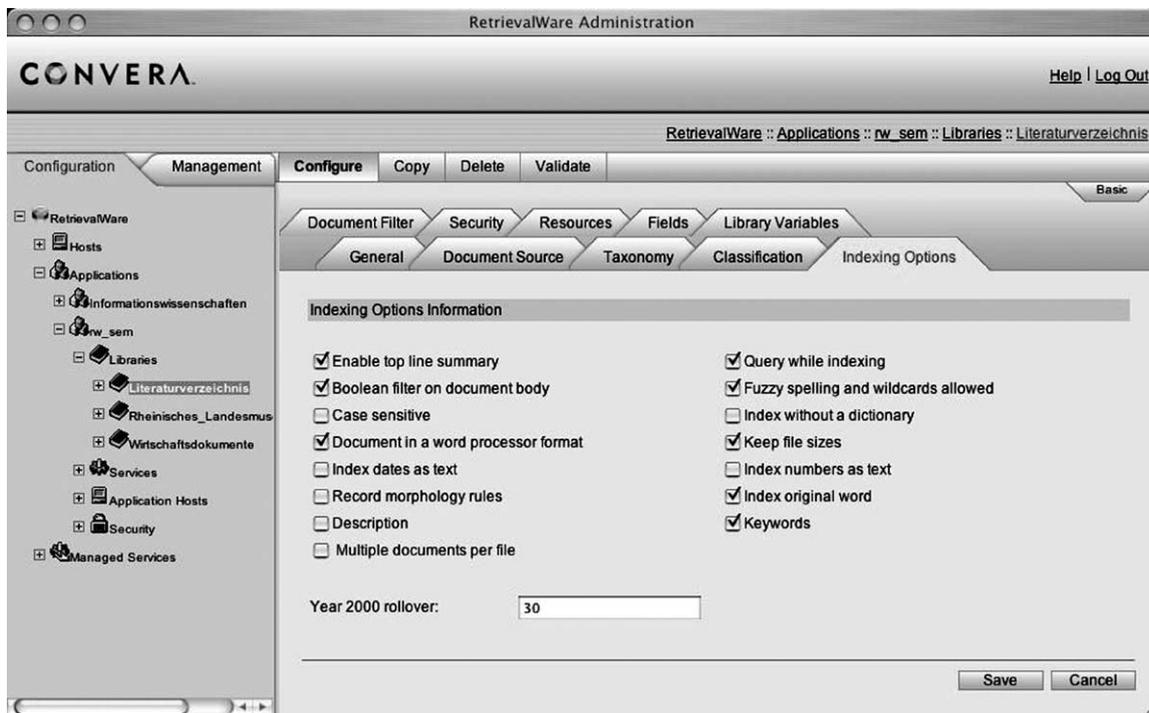


Fig. 4. Indexing Options within the Library Configuration Menu.

5.1. Indexing options

The different indexing options (see Fig. 4) available influence the look of the search result list, the search options for the users, and the indexing process. For further information on indexing options see RetrievalWare System Administration [6, p. 63–64]. Choosing the appropriate options enables a *Boolean Search* for the entire document body, case sensitive search and the more complex possibilities that Convera provides, like *Concept Search* and *Pattern Search*. If the top line summary is chosen, the first lines of a document are displayed on the search result list. Queries can be allowed while the indexing is done, but this slows down the speed of indexing and requires further disk space. The size of the index also increases if dates are indexed as text and if morphology rules are recorded. The latter one means that exact word searches can be performed by enclosing the query word in quotes. In this case, no morphological variants are included. A similar option is “Index original word” which reduces the indexed words to their root forms – this appears to work faster and uses up less index space. If one does not intend to use a library’s *Concept Search*, one can forego the morphological analysis for stemming. Some indexing options must be activated if the library’s documents are not in plain ASCII format or if the stored files consist of multiple documents. The latter also requires some adjustment on the parser so that the boundaries between the various documents in a file can be identified. We presented just a few of Convera RW’s indexing options; several more are available throughout the system.

5.2. Critical consideration

During the indexing of the libraries, error and log files are generated. They are periodically updated until the index is completed, so that the error messages that can be seen directly after the indexing may

show only a small proportion of the existing errors. Furthermore, the system administration can choose between different log file modes, which are advantageous as the default setting yields scarce results. It is advisable to change the mode to avoid misunderstandings. The *Administration User Interface* could certainly use a usability upgrade. The user guidance is complicated, not intuitive and often requires several more clicks than anticipated. For example, if a user opens the management tab, it takes several clicks in the tree structure to get to one of the libraries. While the management tab is open, the configuration can be viewed, but for any changes in the configuration, the user has to change to the configuration tab and go through the whole tree structure all over again. We would have benefited from an explanation of the most common error messages and a suggestion on how to solve the underlying problems. According to Convera's handbook [6] the system should be able to handle 8 bit characters, but obviously there can be problems when documents are uploaded from a Windows computer to a UNIXTM-platform or vice versa, as it happened in the course of our project. Extended characters in several documents were not displayed correctly after the upload to the server and produced multiple error messages during indexing.

6. Recapitulation of the results

It should be pointed out that Convera Retrieval Ware 8.0 truly is a powerful system in the field of knowledge management and information retrieval. With the aid of a database, large numbers of documents can be searched and systematically managed with great efficiency. In excess of simple keyword searching, *Concept-*, *Pattern-* and *Boolean Search* – employed optionally or also in a combinatorial way – represent elaborated ways of information retrieval. Extensive modifications concerning search operations and interface can be arranged in the *Administration User Interface*.

Furthermore, the possibility of integrating a thesaurus into Convera RW must be emphasized positively. The use of a balanced taxonomy is crucial to increasing precision- and recall-values of a retrieval system. Only through semantic and associative connections and relations a thematic categorization becomes possible. Moreover, through the Categorization & Classification Workbench the option of individual cartridge preparation exists, so that domain-particular knowledge can be maintained. The notion of domain-specificity is enormously important, since a data Cartridge that would be able to unite all world knowledge is completely utopian. Classification and categorization function only in restricted knowledge fields. With this method it is possible to specify the content-related terms and definitions of a domain or an enterprise in a relation that cross-links items or separates them from each other. It is important to emphasize that only a group of experts can build a domain-specific cartridge which is able to identify a multitude of concepts and their relationship to one another correctly. Integrating an existing thesaurus is still a daunting task, because the data-structure of the thesaurus may not be ideal for Convera RW.

Along with many advantages of the software, deficits and problems of the system were also encountered. As explained, some points of criticism apply to the area of user-friendliness. Not all individual components of the system are intuitive and easily understandable. This is especially true with regard to the *Search User Interface* and the *Administration User Interface*; however, other components such as the *Cartridge & Classification Editor* could also be improved. Convera Retrieval Ware 8.0 definitely offers a huge amount of interesting and beneficial possibilities for companies in need of "Enterprise Search and Categorization Solutions" [2]. Still, improving these functions to garner greater acceptance from everyday users is an important task for the future. With regard to the search functions, especially the Advanced Search method needs to be explained more thoroughly. Several problematic issues were

further aggravated as a result of the fairly incomplete and misleading documentation. We see the greatest deficiencies in the area of thesaurus construction and implementation as well as in relation to the import functions. The lack of various functions for the simple handling of thesauri complicated the job considerably. When working to integrate the STW, it turned out that a thesaurus of a certain size can only be fully implemented into Convera RW with significant effort. Furthermore, alphabetical sorting of thesauri entries was not always possible.

For large enterprises Convera RetrievalWare, with its numerous possible applications, is certainly attractive. However, the time and manpower necessary to optimally configure and maintain the system should not be underestimated. Only an organisation with a significant cache of unindexed documents and a highly qualified technical staff should undertake an implementation – in that case, the payoff can be considerable. The requirements are certainly formidable and the relation of cost and benefit should be carefully assessed before making the decision to use Convera RetrievalWare.

References

- [1] Altavista: <<http://www.altavista.com>>, visited on 11/17/2005.
- [2] Convera: <<http://www.convera.com>>, visited on 11/17/2005.
- [3] Google: <<http://www.google.com>>, visited on 11/17/2005.
- [4] ISO 5964 – 1985 (E): Documentation: Guidelines for the establishment and development of multilingual thesauri, Geneva, 1985. The full ISO 5964 specification can be accessed at: <<http://www.collectionscanada.ca/iso/tc46sc9/standard/5964e.htm>>.
- [5] J.P. La Hargue, Convera: Categorization & Classification. Standalone test applications, version v1.4.2 (unpublished).
- [6] RetrievalWare System Administration, Convera RetrievalWare, version 8.0.3 updates, 9/7/2004.
- [7] STW – Standard-Thesaurus Wirtschaft, ed. by: Bibliothek für Weltwirtschaft – Deutsche Zentralbibliothek für Wirtschaftswissenschaften (Kiel), GBI (Muenchen), HWWA (Hamburg), Ifo-Institut (Muenchen), 1998.
- [8] WordNet 2.1: <<http://wordnet.princeton.edu/perl/webwn2.1>>, visited on 11/17/2005.