An innovative approach to access unstructured enterprise information and to improve availability of enterprise knowledge

Author Name(s): Stephan Wurst, BALance Technology Consulting GmbH, Bremen, Germany, stephan.wurst@bal.eu, Markus Lehne, markus.lehne@bal.eu

Abstract

Knowledge is the most valuable resource of a competitive company. To make this knowledge available to its employees and to keep it alive is one of the most challenging tasks in modern business.

Shipbuilding has some characteristics that require special measures in order to preserve knowledge and provide it to the employees. Ships are highly complex one-of-a-kind products. Errors in design and production cause high costs and might not even be corrected. Therefore it is crucial for the competitiveness of a shipyard to avoid such errors by documenting the solutions for the engineers This type of information is often unstructured which makes it even more difficult to capture it in a knowledge system. It might also occur in sources that are not covered by the available search mechanisms such as e-mails and their attachments. Secondly knowledge is often bound to particular persons, which means that there is a high probability of information loss once these experts retire or leave the company.

To support the processes of gathering and disseminating knowledge, an innovative knowledge management approach has been developed which supports "knowledge workers" by improving the speed to get their individual questions answered with a tailor-made set of documents structured by a company-specific category tree. It offers an interactive, living knowledge base with support for preparing electronic training material with minimal effort.

The approach described in this paper is based upon a company-specific category model for matching the knowledge and production experience onto the product and process structure of the company.

Keywords:

Knowledge management; categorization; document analysis; access to unstructured knowledge; feedback; automatic document processing; Multilanguage support

1. Introduction

To support the process of gathering and disseminating knowledge for easier access by knowledge users and providers, an innovative approach has been developed. It mainly supports four aspects:

- Provisional targeted access for all knowledge users to unstructured enterprise information
- Optimised support for the "knowledge workers" in a company by significantly increasing the speed to get their questions answered by the company's information base even when the documents are available in different languages.
- A consistent up to date and thus living knowledge base that is continuously growing without requiring extra effort, by efficiently supporting online discussion.
- An efficient infrastructure for structuring the knowledge, e.g. for preparing electronic training material or best practise handbooks.

The major challenge when implementing a new process or introducing a new (software) tool is to achieve its acceptance by generating immediate benefit visible for the people involved. In case the new process does provide only benefits without requiring any extra effort the

major stumble stone for a successful introduction is avoided. Therefore the main philosophy for our approach is

- automated processes where possible and sensible
- strong visual support especially when evaluating the search results
- orientation towards known company procedures
- no changes to existing data storage structures or procedures

It is a major objective to support both knowledge providers and users by explicitly addressing their specific needs. The document import process is carried out automatically and search results are presented through a hyperbolic tree that shows the relationships between knowledge elements. Knowledge consolidation is supported by easy-to-use software and a discussion forum for feedback collection.

Central part of the concept is a company-specific category model for matching the knowledge content onto the product and process structure of the company, to allow automatic categorisation of all imported documents.

2. A Challenge and its Solution

When trying to access company knowledge, some typical aspects have to be considered. The users want to

- **Find the information they need**.

  Therefore a powerful search engine is needed that supports the user in generating the queries in an easy yet powerful way. It has to be ensured that all relevant information is found.

- **Trust the information.**

  It has to be ensured that the information stored in the database is accepted by the knowledge experts either as general information or as approved company knowledge. If this cannot be guaranteed (e.g. for information retrieved from the Internet) such results have to be clearly marked.

- **Make utilisation of the information as easy as possible.**

  Presentation of results should on one hand support an easy reading of the information found. On the other hand relationships between documents and recommended further reading should be made available intuitively. To avoid confusion of the users presentation must not overload the reader with information but should present the results successively.

- **Use one integrated solution for all types of users.**

  Independently of the skills and the type of work the users do they should be provided with a single user interface. The only difference should be that unavailable functions (like changing settings when not having the authorisation to do so) should not be visible.

- **Make the information available to those needing it.**

  All knowledge users should have access to the system. It should especially be not limited to engineers only. Restricting the use to the user's domains is recommended though, especially when modification or discussion of the content is involved.

- **Make sure the internal experts have consolidated the information into knowledge.**

  The knowledge base contains content crucial for the company's work. Thus it is absolutely essential that only information approved by company experts is accessible. The only exception is content clearly marked as preliminary or unapproved.

- **Support discussion of the knowledge (keep it alive).**

  Even though not everybody is allowed to enter data into the knowledge base, all users should be enabled to comment the company knowledge they have access to. By following this approach it can be guaranteed that the knowledge reflects the current company processes without hindering the evolution of these processes by considering experiences gathered through knowledge use.

- **Avoid extra effort.**

  The end user does not have to do anything to get the documents into the system. He may generate and discuss new knowledge when using the software but even without active participation in the knowledge creation process, the environment can be used in an effective way.

For the administrator, three tasks remain:

1. Define the category tree
2. Train the tree
3. Select directories to be imported into the knowledge base

A potential solution therefore has to consider the requirements of the end users. A knowledge user wants to find all relevant information for a given problem and needs a visualisation of the search results that supports him in identifying the most suitable subset of the documents found.

The knowledge provider is interested in a fully automatic import process. No "auxiliary information" (like meta data, keywords, etc.) should be added manually. Instead a monitoring facility that recognises document changes and imports the changed files accordingly should be available on every server relevant for the system. Furthermore, the target group needs a discussion environment in order to get feedback for the created knowledge which keeps the knowledge alive.

The administrators of the document and knowledge management do not want to install just another database that again contains all the documents. They prefer software that interfaces the existing information sources and only adds the new functionality thus reducing redundancies.

3. The BAL.KMAN Approach

After having analysed the "knowledge process" in different companies, a best practice common process has been identified:

1. **Searching the company information base**, going through archives, performing an Internet search, searching the company network to answer questions like: What does the company already know? Is it necessary to complement this knowledge? Where is the information? Can the information be trusted? Is it up to date? Which are the company experts? Most of these questions can not be answered from a pure document

search but by an additional discussion with the company experts. Therefore the search has to be seen as the starting point of gathering the required information. The search process it self should be oriented at commonly used search engines such as Google. Using a similar syntax increases the acceptance since the users can start using the new tool immediately. Results of a search can be shown as a conventional list but to make use of extended features such as categories or automatically found entities, a hyperbolic tree view is more appropriate. Such a tree can show relationships between documents, thus guiding the user through the result set and enabling him to find documents that would not necessarily have been regarded as relevant without the knowledge about the connection that is being shown by the tree.

2. **Compiling the search result and creating a "summary" from the extracted knowledge,** is a typical way of working. In particular: collecting the result of the study, sort it, analyse it, digest it, compile it. So common these steps are so different are the ways of doing them. Some people build piles of papers, others create files on a computer, others pin the findings on a wall and draw connection lines between them. The concept aims at supporting the user in this compilation process by transferring information into wiki-like "articles".

3. **Consolidating the summary into an "article" through discussions with expert colleagues,** is the most logical way because discussing the compiled result with your colleague experts helps verifying whether an aspect or a piece of information has been missed and thus proofs the value of the article. Consolidating the findings and documenting the outcome of the study in form of a memo, or presentation will thus help not only the searcher but potentially also the rest of the company. Such a publication has to be marked as "personal opinion" in order to not accidentally use it as reference document, but it can now be evaluated by colleagues and experts in the company.

4. **Presenting the "articles" to the entire staff as reference Know How.** This Step is neither well established in today's companies nor is it easy to refer to "one best way". Some companies use internal papers to spread the knowledge, others prefer workshops or company websites to inform the staff. In any case the outcome of this step should either be a document that represents the company knowledge in the domain addressed or the conclusion that the document has to be withdrawn.

5. **Share /disseminate/educate - Using the articles to build lectures or best practise handbooks for internal training.** Once you have a consolidated documentation of a certain topic this documentation will be made public to those in the company interested in the topic. This would enable for both providing the information in a database for individual use but also to utilise it in form of classroom lectures to actively train people on a particular topic. Major advantage of adding this step is the structuring of the knowledge thus supporting the reader in identifying related documents and guiding him in the correct order of reading them.

6. **Combining individual team member knowledge and company know how to create new input for the company information base,** is the ultimate step in this

process. Companies which successfully have implemented these steps contribute to the company culture with a good support for a lively innovation atmosphere.

BAL.KMAN is the BALance Knowledge Management package that reflects the requirements mentioned above and implements a solution to help knowledge workers in using the knowledge as efficiently as possible. Three modules exist:

- **BAL.ASK**
  - Store and retrieve documents
  - Search engine
  - Result visualisation
- **BAL.PEDIA**
  - Discussion of knowledge
  - Identification of company reference knowledge
- **BAL.ELECT**
  - eLecture and handbook editor
  - Preparation of training material and best practise handbooks, structuring of knowledge

Together they realise the "knowledge circle" as represented in figure 1. It consists of the six steps mentioned above (which are common practise in a "knowledge workers life") and ensures that all company knowledge is captured and made available in order to not miss anything important when carrying out the daily work.

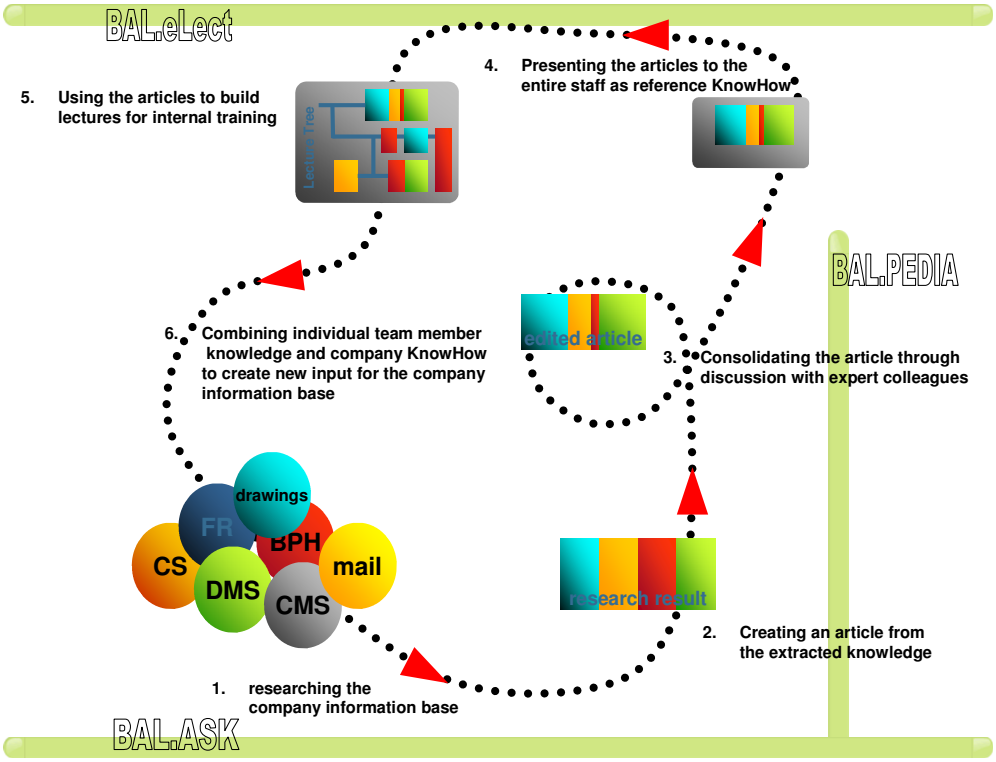

Figure 1: Knowledge circle

4. The Knowledge Circle in Detail

It is assumed that the relevant knowledge is already available in digital format. Sources could be file servers, content and document management systems, company standards, best practise handbooks, mail servers etc. By means of monitoring tools these documents are transferred into the BAL.ASK component of the BAL.KMAN system. Here they are analysed, categorised, entities are detected and a summary is prepared. All this is done fully automatically by the system based on linguistic and statistic algorithms. After this procedure they are available to be searched and displayed by the users.

They can now submit queries to the information base and get a result that will be viewed either as a hyperbolic tree or as a list. While the list is mainly meant for cases where a certain document is searched for the hyperbolic tree is the standard way of looking at the result. Here the user can identify related documents and all categories or entities that are related to a particular file. By further unfolding the tree and browsing through it, documents can be found and opened that might not have been the first choice but also contain useful information. Exploring the information base using the hyperbolic tree might there open the view to aspects of a given problem that might have not even be considered initially.

As a second step qualified users (the so-called “knowledge experts”) generate articles that conclude the knowledge from documents found, external sources and newly written text. These articles are then made available by the BAL.PEDIA component and may also be searched via BAL.ASK. In this stage the articles are marked as preliminary and can be changed by authorised users.

Now the third step starts, which involves the consolidation of the knowledge by discussing the articles with users and other experts. During this process, the article might be edited iteratively until the requirements are fulfilled in an acceptable way. This actual editing is exclusively carried out by the experts while contributions to the discussions can be provided by all system users.

Next the article will be marked as approved knowledge in step 4. Such articles might neither be edited nor deleted any longer. The only possibility to alter such knowledge is to unprotect it again which is only allowed to a number of privileged users, the “knowledge supervisors”. Only if they agree, approved company knowledge will be marked as outdated or to be revised. Approved articles represent the company knowledge which the users can rely on.

In order to structure the knowledge the 5th step has been introduced. Here the existing articles will be combined to groups which cover lectures or handbooks to introduce work processes or comprehensive concepts to new employees or to people who have been assigned to new responsibilities. Drag and drop technology allows to combine lectures and articles to form new lectures which finally results in a tree of related knowledge elements.

Finally, the information about lectures, handbooks, articles and discussions is fed back into the company information base.

5. Making Knowledge Management more User-Friendly

5.1 Increasing user acceptance

One major problem when introducing a knowledge management system is the user acceptance. Typically, the users are not interested in maintaining the knowledge base or document base but in getting the information they need in order to get their daily work done. Therefore the most important issue to address is to get the documents into the system with an as small number of user interactions as possible, preferably none at all. Furthermore the

functionality must allow easy access to the database content regardless of format and language of the information in question. The knowledge management approach described in this paper therefore includes a number of features that aim at increasing the user acceptance. Some of them are unique such as the comprehensive multi-language support, some others, e.g. result presentation as a tree, have been implemented in other projects as well. However, the combination of all the functionalities has not been realised in a single environment so far.

5.2 Collecting information

Import of documents is carried out automatically. Document sources can be file servers, e-mail server, document management systems, content management systems etc. In case of an e-mail server, not only the content of the message as such is analysed but also the attachment.

The directory monitor used for managing the import process can be installed on any client machine in the company network and monitors selected directories then. Meta data like owner name, project and organisation can be assigned to each document in a certain directory. If access rights are provided they can be set here as well. After registration of a directory it is supervised and the change requests are sent to the BAL.KMAN server.

Since the assignment of attachments and messages is retained it is possible to navigate through the content of the mail server using the hyperbolic tree functionality.

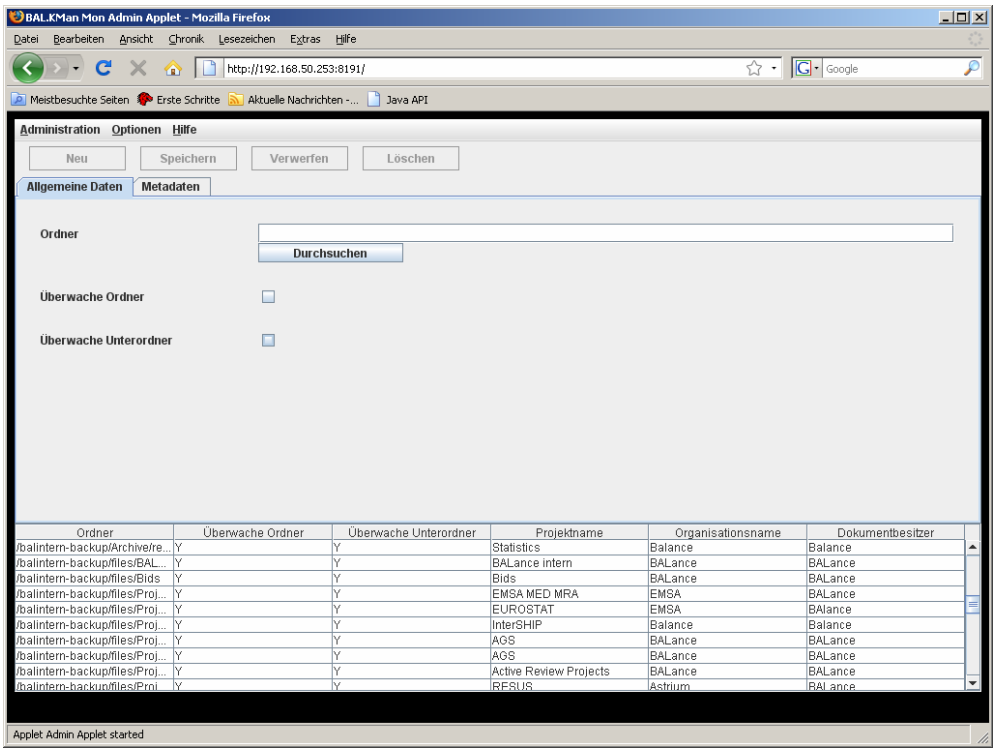

Figure 2: Monitoring a file server

Most of the commonly used file formats are supported. It is possible to import office documents such as Word, Excel or Power-Point files. Furthermore, PDF, HTML or text files can be read. In total approx. 50 file formats are currently supported. However, it should be noted that the import quality depends on the file format. The more structural information a document contains the better will be the result of the analysis by the linguistic components.

The language of a document is recognised automatically which is an important feature for the search but also for the following analysis of the file content. One step is the automatic categorisation. Depending on the company-specific category model (taxonomy), each

document is assigned to a number of categories. This process involves some manual interaction by the knowledge manager beforehand as the model has to be defined and trained first in order to match the company document base. After this initial step the system automatically assigns the documents to the categories. Therefore the knowledge user does not need to perform any manual task during import.

In the next step keywords (or entities) are extracted. Although the rule base to identify these entities is pre-defined it can be adapted to the specific requirements of an organisation. Typical expressions found are names, dates, locations, products etc.

Finally, an abstract is generated. It contains a pre-defined number of sentences (5-10) that are automatically identified. This summary gives a first impression of the content when browsing through the result of a search.

5.3 Searching and retrieving

A wide-spread problem when searching through a document base is that large files might match the query but still contain a lot of irrelevant information in the given context. In order to overcome this problem, such texts are automatically split on chapter level. A soon as a document exceeds a specific threshold (10 pages per default) it is split. The pieces are analysed again. If a chapter is still larger then the threshold it is further divided into pieces. These text blocks are then imported as single documents. However, the connection to the originating file remains in the database enabling the user to still get the full file if needed.

A powerful search engine has been developed which allows searching the entire information base for documents, e-mails and their attachments, articles and related discussions and finally best practise handbooks. The actual query syntax is quite similar to well-known Internet search environments. Boolean operators can be used as well as wildcards. Some further functionality is available though. The search can be restricted to documents belonging to a certain category or entity, written in specific languages or having a pre-defined file type.

On the other hand the search can be extended by allowing spelling deviations and by searching for documents similar to the one that has been found. All these features can be used on their own or in combination.

During the search, the terms can be translated into different languages semi-automatically. This service is to be initialised by the actual system users. Every time they are missing a translation it can be added manually and will then be useable by every other person using the software. Even different translations for a single phrase are possible. In this case, the user can selected one or more meanings for a given search term. In order to avoid inadequate translations they are monitored. The one that is selected most will be presented first to other searchers.

The system supports most common languages. A total number of more than 30 languages can be added to the system if needed.

After the query has been processed, the result can be visualised as a list or graphically as hyperbolic tree where the search results appear as leaves while categories and entities are shown as the branches. This tree can either be based upon the category model or on the entity tree (keyword tree). It is possible to navigate through the tree by open further nodes dynamically to show a document's entities/categories or vice versa. The tree can visualise relations between documents, e.g. other documents of the same category etc. Thus it allows navigation through the information base in an intuitive way. Hyperbolic trees also allow reducing the number of relevant search results quickly by simply restricting the analysis to only a certain part of the tree, hiding all documents that fit to the query but belong to a

different domain.

When watching the actual documents, different file formats are available. Typically, the HTML view is used which highlights the search terms and allows the direct export into an article. Secondly, all documents are also available as PDF file for printing them in their original layout. Finally, each document can also be downloaded in its original format if the user rights are set accordingly.

5.4 Creating new knowledge

Investigating a certain topic might lead the searcher to information in documents which can directly be used as input for article creation. Without any further processing they can be converted into an article and then be edited directly. The articles serve here as a collector for all the bits and pieces of information which are considered relevant for the study, and can gradually develop from “first notes” to an excerpt on the topic to a consolidated paper. After article creation it will automatically be inserted into the database with the same procedure that is used for other documents as well. This means that article and documents are fully integrated in the database and can be searched without using different tools. Independently it is possible to create articles that are not based on documents but are created from scratch.

To support collaboration between employees the discussion of articles is also integrated into the BAL.PEDIA module. By simply clicking into the discussion area a new contribution is generated and then automatically inserted into the database. From the searcher’s point of view there is no difference between articles, discussions and documents. Everything is found as long as it matches the query.

Finally, articles can be grouped into handbooks. A handbook comprises an introductory text in the same format and with the same functionalities as an article. Attached are other handbooks or articles which are presented to the user as a single piece of information structured by an automatically generated table of content.

Handbook creation and editing is done by a software module similar to the article editor. As an additional part it contains the handbook combination part which allows to drag and drop chapters or articles into the current handbook. In order to avoid handbooks that contain preliminary information only articles that have been acknowledged by a knowledge manager may be inserted into a handbook. This process is called “article protection”.

It is important to note that by using BAL.KMAN the user does not change documents in the information base. Although it is possible to edit articles and handbooks, the source documents remain unchanged. In case a document is downloaded in its original format, edited and stored on the local hard disk, this will not affect the BAL.KMAN database. Only explicit changing of the original file in a monitored folder or copying the edited file into a monitored folder will lead to a new database entry in the BAL.KMAN system.

6. System Architecture

The system architecture of BAL.KMAN can be divided into three major components. The document base consists of already available sources that provide information objects. Possible software systems are file servers, document management systems, content management systems and e-mail servers. Additionally, the system is also able to communicate with PDM and expert systems, such as Quaestor which allows the realisation of additional features like life cycle cost calculation. These systems do not actually belong to BAL.KMAN but are accessed by the core system via APIs (Application Programmer Interfaces).

BAL.KMAN server and client actually realise the BAL.KMAN functionality. The server implements the import and analysis features as well as the search and retrieval engine. During import of a document the analysis modules are needed. First the document is received by the import service. It sends the document to the converter which extracts a PDF, an HTML and a text only version of the original documents. Larger documents are split first. In the next step, the document language is detected. If the language is supported by BAL.KMAN, summary, categories and entities are extracted. All this information is then stored in the document index. When access rights have been submitted as well, the information is stored in the administration database.

The second process supported by the BAL.KMAN server is the search and retrieval for the content. A query is sent to the search engine which retrieves the document matching the expression while considering search restrictions with respect to categories, languages, document types, etc. Before returning the list of found items, accompanying information is added and the administration database is consulted in order to check whether the documents might be accessed by the inquirer. Finally, the list of documents is sent back to the client.

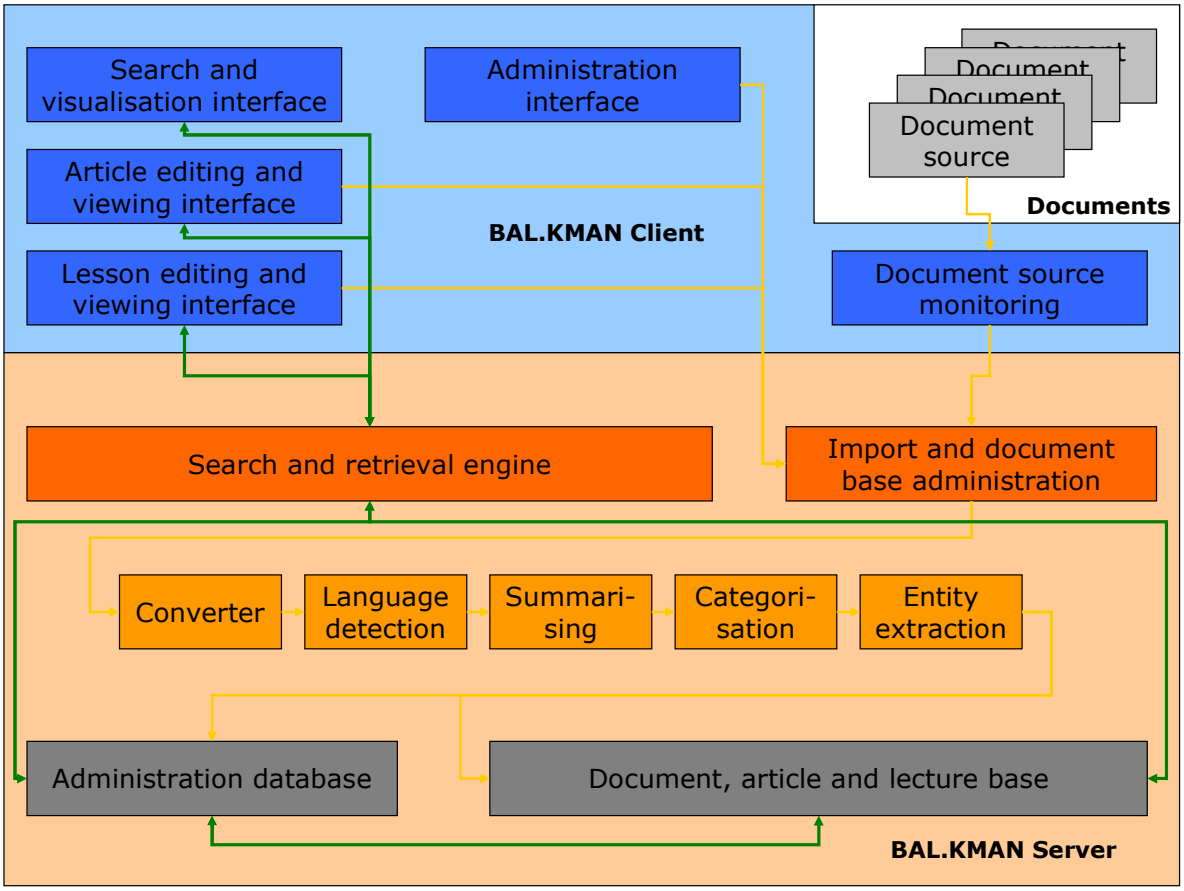

Figure 3: BAL.KMAN system architecture

The client components implement the user interface to

- Search for documents
- View the search result
- Generate articles and handbooks
- Maintain the database

All this functionality is realised by web browser based applications. An additional stand-alone module monitors directories on file server and e-mail servers and transfers modified and new documents to the server.

7. User Interface

The typical entry point for submitting a query to the BAL.KMAN system is the main search page. On the left hand side the category tree is shown and provides means to restrict the search to a subset of all topics. On the right hand side the query is entered and can be further configured by providing translations for the expressions, specifying the type of documents to be searched for, the presentation of the result and the precision of the query. Most of the settings are optional. The only mandatory field is the actual query. The entire user interface is available in different languages. In this document, German and English examples are shown.

Managing of the translation mechanism is done by the users. Whenever they find out that for a given word no translation is available it can be directly added. For each word an arbitrary number of translations can be inserted. As soon as another user types in the search term the translation appears automatically. If more than one translation is available the one selected most will be proposed as default. Alternatively, several translations can be used simultaneously, e.g. “ship” and “vessel” for “Schiff”. In this case the selected search would be (“ship” OR “vessel”) OR “Schiff” which means that a document is found when it contains any of the words.

Figure 4: Main search window

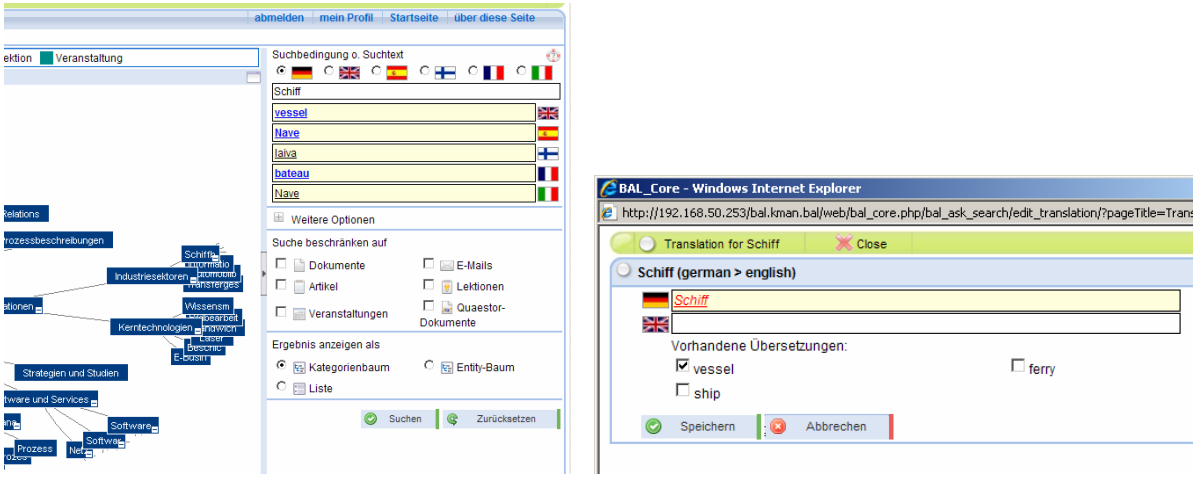

Figure 5: Semi-automatic translation of search expressions

The search result can be either visualised as a list or as a hyperbolic tree. This tree does not only show the assignment of documents to categories or entities but also the relationships of documents to each other. Different types of results are coded by colour. E-Mails are shown together with their attachments. Handbooks are connected to their articles which may be connected to discussions. It is possible to further browse through the tree by opening the items connected with a node. Only the system memory sets a boundary to the amount of nodes that may be shown in the tree.

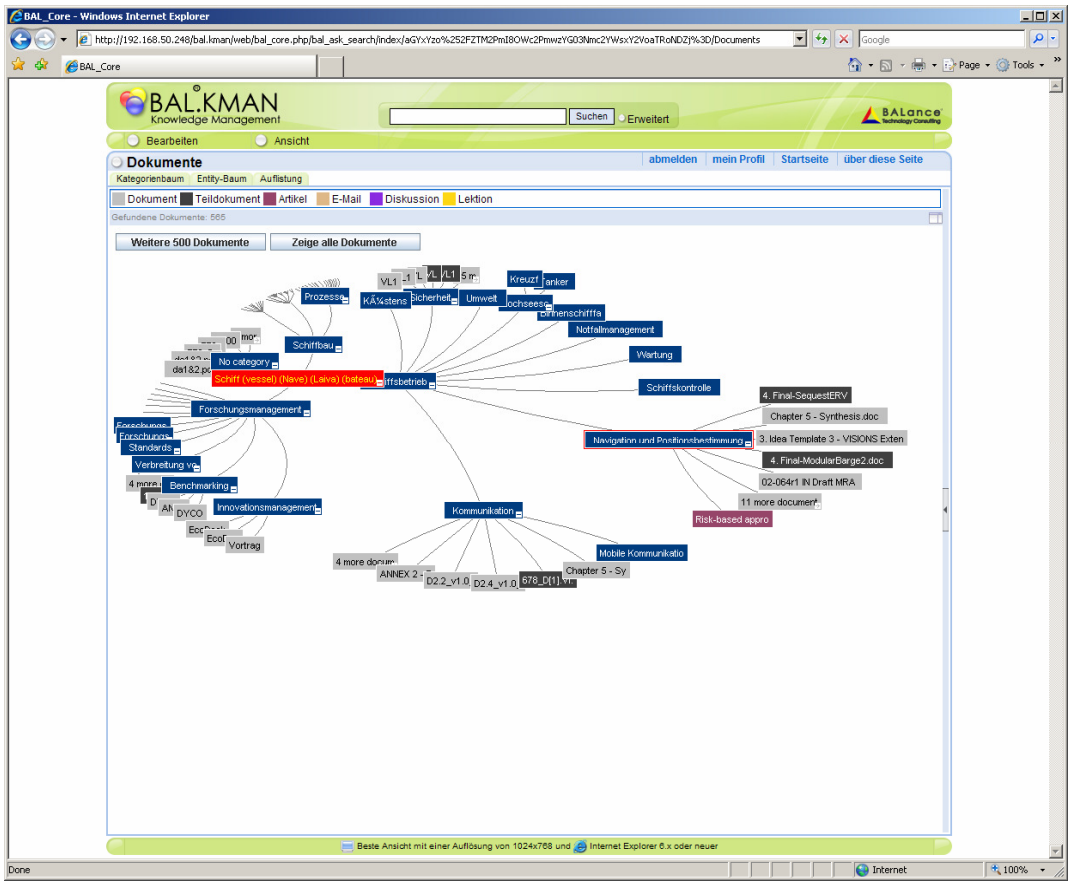

Figure 6: Result view as hyperbolic tree

The article view lists all available articles in alphabetical order together with author, language and date of the last change. Approved articles are marked with an asterisk and cannot be modified any more. Only these articles can be used by the BAL.ELECT module to create handbooks. By clicking on an article it can be viewed.

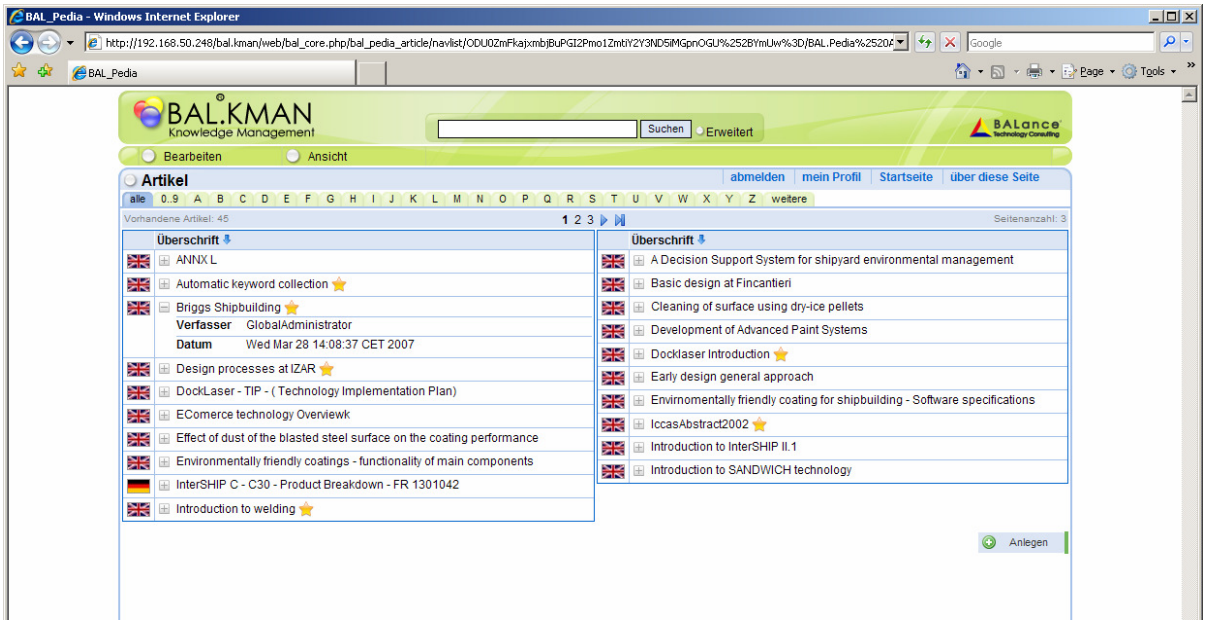

Figure 7: List of articles

The article view shows the actual article content and additionally the discussion contributions. If the user has the right to edit or protect the article, these functions can be accessed in this window as well.

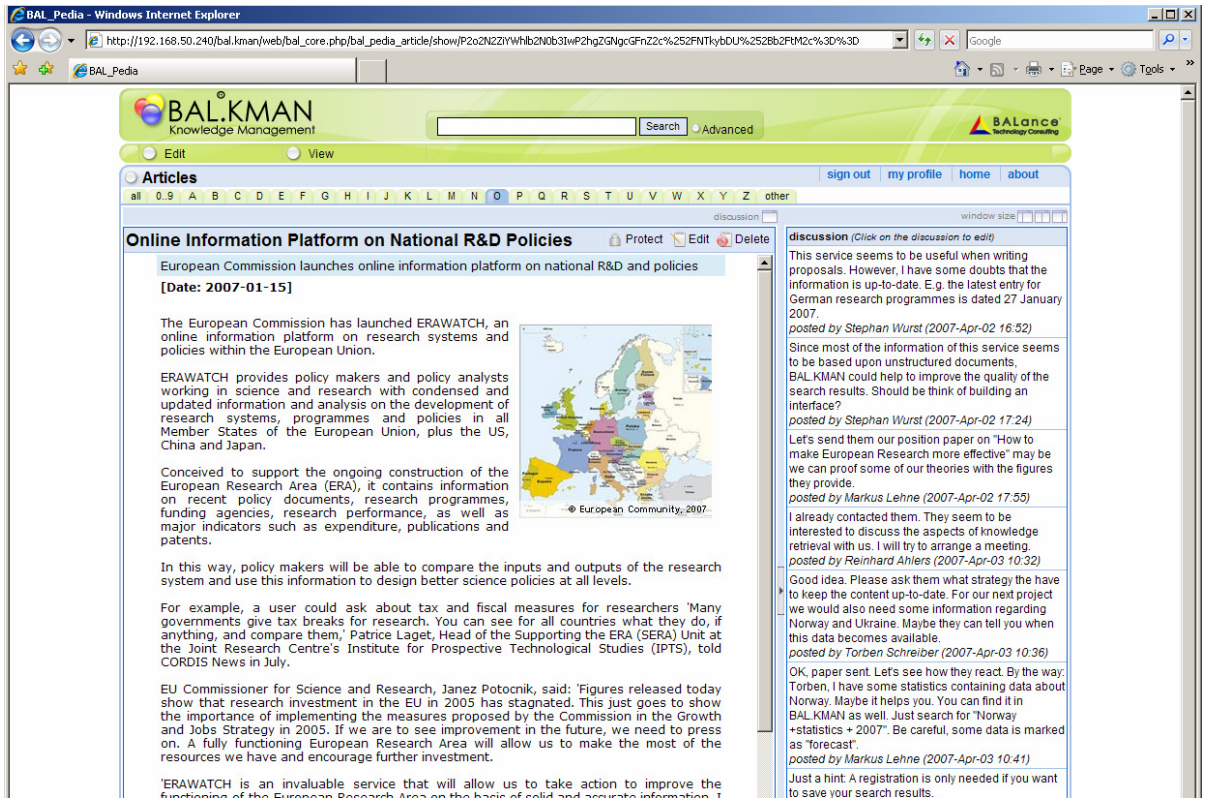

Figure 8: Article view

A handbook is combined of text, articles and chapter (“sub handbooks”). The BAL.ELECT module allows the combination of these components by drag-and-drop technology. Approved articles and handbooks can be dragged into the handbook and are then immediately available for other users.

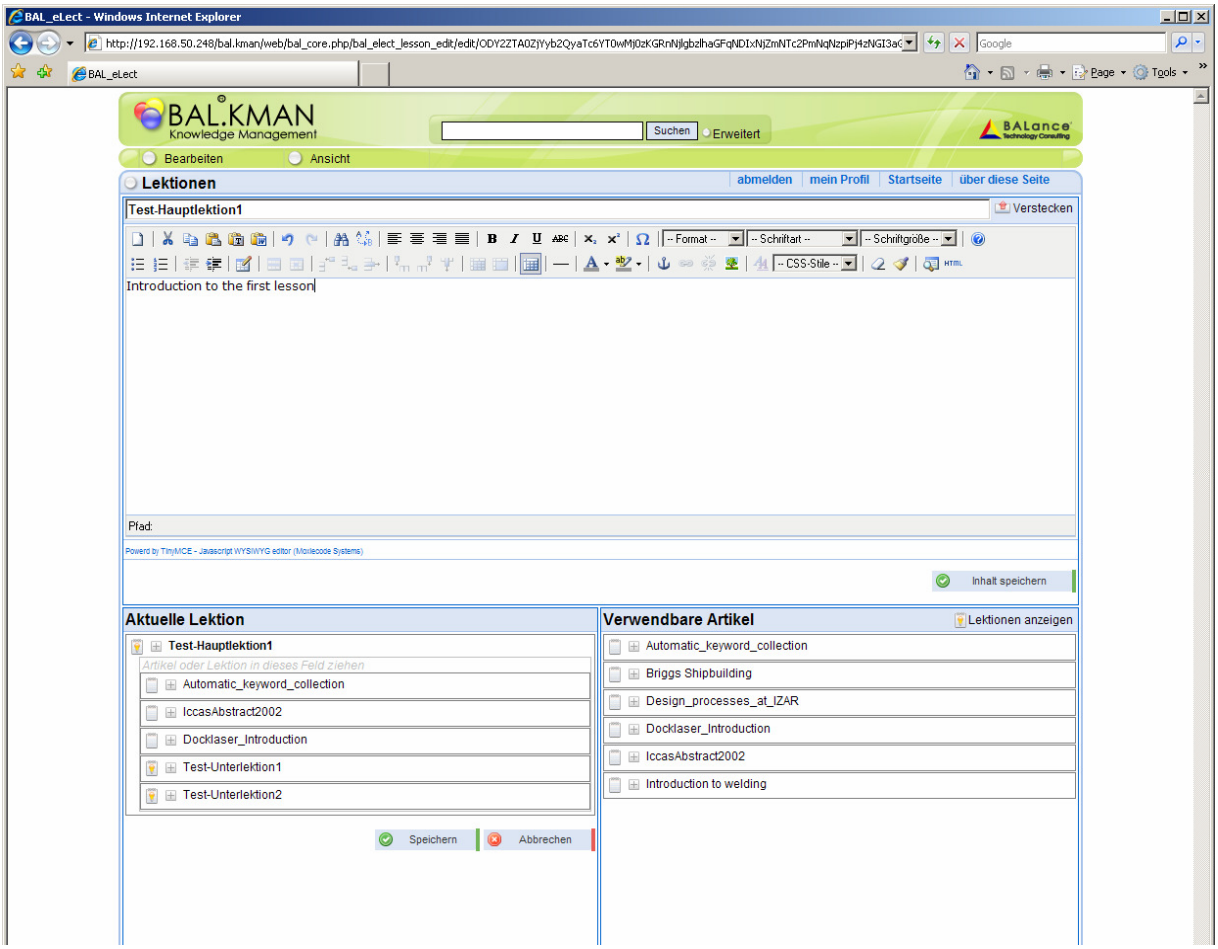

Figure 9: Editing a handbook

8. Main advantages of introducing BAL.KMAN

All users benefit from finding information in an unstructured environment quickly according to the tailor made ontology tree. Especially employees working for enterprises operating in multi-language environments can reduce the effort of finding relevant documents via the automatic translation of queries.

Users have the possibility to learn about the company processes fast by reading the articles that contain the reference knowledge. Since the system motivates the users to discuss existing and to contribute their own knowledge it is ensured that the value of the content for each single user grows with every new article written. Supporting new employees in getting information about internal processes quickly makes it easier for them to integrate into their work environment.

Summarised it can be said that using BAL.KMAN decreases the time needed to find important information while increasing the willingness to provide own knowledge to the colleagues.

For the management additional advantages are apparent: The statistic on search terms and results found shows the gaps in the knowledge base which can directly lead to effort for improving the company knowledge thus increasing the skills of the employees and improving competitiveness in knowledge-intensive domains.

The enterprise knowledge, even from employees who have left, is condensed in a structured database making supporting training measures for new employees as well as for the experienced staff. Discussion about processes shows weaknesses in the way of working, giving the opportunity for corrective measures even before actual problems arise.

In general installation of BAL.KMAN helps the enterprise to organise itself and leads to a base of lively and thus up-to-date information.

9. Conclusion

BAL.ASK, BAL.PEDIA and BAL.ELECT offer an innovative approach to collect and use company knowledge. Major emphasis is put on automatic document import and an easy to use yet powerful user interface.

The BAL.ASK module is a powerful solution to store and retrieve documents. BAL.PEDIA supports the discussion and consolidation process to extend available company knowledge. BAL.ELECT supports the creation of eLectures or eBooks and therefore the structuring of knowledge.

Together they can help each organisation that owns a big amount of unstructured knowledge to sort it, structure it and make it available to everybody who needs it.