Ontology-based Knowledge Management for SMEs

Markus Schwinn, Norbert Kuhn, Stefan Richter

Institut für Softwaresysteme in Wirtschaft, Umwelt und Verwaltung
Fachhochschule Trier, Umwelt-Campus Birkenfeld,
Campusallee, 55761 Birkenfeld
{m.schwinn,n.kuhn,s.richter}@umwelt-campus.de

ABSTRACT
The OnToBau research project\(^1\) aims to provide a way to classify, archive and effectively use their knowledge with the assistance of an ontology-based knowledge archive for small and medium companies from construction industry. This archive is intended to pro-actively provide users with information to assist them in their daily business process handling.

In this paper we describe the architecture of the OnToBau system. The system has four main parts. The preprocessing-modules prepare the resources that should be stored in the knowledge archive for the enclosed inference system. The inference system is the core component and extracts the information from the preprocessed resources. Ontologies provide the necessary domain knowledge. In order to exploit the available knowledge, a personal agent monitors the current activities of the user and tries to infer the intention from his behaviors. At certain points it automatically offers the user helpful information. Again ontologies are used to represent information about the business processes. In addition, the user has the option to search for information in the archive through the graphical user interface. In this paper we focus on the description of the first main part of the architecture: to easily add knowledge to the knowledge base and to allow a general processing of different resources within a company (E-Mails, calendar entries, papers etc.), an XML-based representation language was developed, namely the OnToBau representation language (ORL). Converters are used to transform the resources into ORL.

Keywords: Ontology, Knowledge Management for construction industry, intelligent support for processes

1. INTRODUCTION
The growing importance of computers in the 70s and the associated opportunity to disseminate information in digital form is considered as the turning point to today's information age. As with any other technological revolution in history, the new information technologies spread within two decades all over the planet [1]. For the first time, it was possible to produce, to copy and to archive information in a simple way.

In the early 90s another technological milestone was passed, that moved the world to a new era of globalization [2]. The Internet offered the opportunity to access and produce information on an increasing number of websites.

\(^1\)This work is supported by the German Ministry of Research and Technology under grant FKZ 1731X09

The largest flow of information in the Internet is produced by e-mails. A study of the Radicati Group in 2009 stated that in the same year about 247 billion e-mails per day would be sent, and that this number would be doubled by the year 2013. However, approximately 81% of those e-mails can be considered spam [3][4]. This information overload will be a challenge for many companies in coming years. This is also covered by the same study: it is estimated that companies with 1,000 or more employees have to invest 1.8 million dollars per year in the processing of e-mails and spam [4].

However, not only the flood of information from the growing e-mail traffic will challenge the companies in the near future. A study of the Gartner Inc. in 2002 amongst around 300 companies showed that 96% suffer an information overload. To counteract the information overload, companies will have to spend 30 billion US-Dollars within the next years [5]. A similar investigation of the Basex Inc. in 2008 stated, that the consequences of information overload produce costs of about 900 billion US-Dollars a year for the economy of the United States, caused by reduced productivity of the employees in knowledge intensive processes, which have to spend about 25% of their daily work with searching for information [6].

Interviews with our project partners from the construction industry provided similar results. In particular, the quotation preparation process has been described as a knowledge-intensive process. The reason is that different resources are used to produce a quotation. Basically, a product database is used. If the product is not in the database, the employee searches in product catalogs, previous invoices or web pages. In addition, business documents are archived in paper form in many small companies, as document management systems are often not tailored to their needs. So searching for information is often a very time consuming task [9].

In this paper we describe an architecture, which should pro-actively provide these small companies with relevant information stored in their knowledge archive to assist them in their knowledge intensive business processes. Because of the enormous variety of different business processes our system focuses on processes in the construction domain (e.g., the quotation preparation process). Surely the results will be adaptable for other domains with similar processes.

The knowledge base is built up from different resources a company has to deal with (e.g. invoices, product catalogs etc.). A personal agent should monitor the behavior of the user on the computer to infer the user’s intention and automatically provide him with useful information.

The remainder of the paper is organized as follows. In Section 2 we give an overview of related work. The overall system architecture is presented in Section 3, with an overview of the main components and the possible interaction abilities an end-user will have. In particular we discuss the OnToBau Representation...
Language (ORL) in Section 4. All relevant documents are presented in this XML-based description language. We will also present the approach to convert different documents into the ORL. Finally, we end up with the conclusion and outline some future work.

2. RELATED WORK

There seems to be no other projects that specifically consider a knowledge-based process support in construction industry. However, the project DYONIPOS, which has a strong similarity to the aims of the OnToBau research project as described above, tries to optimize processes in public administration facilities by providing proactively the available knowledge to the employees [7]. DYONIPOS has adopted a strict process-oriented approach that moves the focus to the business processes [8]. This approach is appropriate in an environment with highly structured processes, like in public administration. However, in an environment like the construction industry most of the processes are semi-structured. In our approach we focus on the documents and the knowledge contained within them. Therefore, providing knowledge is more likely to be tied to the user’s behavior than to rigid processes [9].

Chung et al. describe in [10] the Task Based Process Management (TBPM) project that aims to support the management of change in business organizations with the help of intelligent task management and coordination technologies. The area of new product development (NPD) within the chemical industry was chosen as the application focus for generating and testing ideas. The project extends the scope of current workflow systems by using models representing knowledge of the application domain and organizational context. Both, knowledge about the domain and the processes, are represented in ontologies. The ontologies are used in several ways to support the processes, but in contrast to our approach, they don’t provide the user with relevant information pro-actively. For example, they are used to determine which process model is best to perform a given task or to pass information along the process chain to match information requirements with information generated by other tasks.

In [11] the Information System Research Group Swiss Life presents a knowledge-based system for supporting office work, namely the EULE2-system that aims to provide a user with a maximum guidance in performing office tasks he may not be familiar with, especially tasks where many laws and regulations influence on how a certain task is to be performed properly. Office tasks are visualized as a graph. Its nodes represent states in which the user must perform certain actions. Some of the actions (like generating letters) are performed by EULE2 (possibly delegating it to another application system), the others are done by the office worker, telling EULE2 when they have been completed. Only when all actions associated with a state have been performed the user can change to a subsequent state. In guiding the user through an office task EULE2 supplies him with exactly the knowledge that he needs at a certain moment, namely what to do next and why (the latter happens only if he is interested to know). To achieve this functionality EULE2 has knowledge about office tasks and all the laws and regulations that must be observed by the office tasks. That knowledge is represented declaratively by first order deduction rules. Thus, EULE2 is an example of a just-in-time knowledge provision system [11]. In contrast to our approach, EULE2 is in some sense very similar to a workflow management system that supplies one office worker with relevant knowledge on laws and regulations to fulfill an office task.

3. OVERVIEW OF THE ONTOBAU-ARCHITECTURE

There has been much literature about knowledge management systems (KMS) within large enterprises and little information available on KMS within SMEs [13]. According to [13] SMEs have special requirements on KMS. Interviews with our project partners led to the same result. The managers in SMEs are in most cases the owners. The result is, that the decision-making process is shorter than in larger companies. They show a flat and less complex structure, with fewer layers of management [14]. Processes are often not as strongly structured as in larger enterprises and knowledge is distributed at various points in the company (file folder, product catalogs, databases)[9]. A smaller number of people within a company is usually united by common beliefs and values, resulting in shorter and often less strategic ways for making decisions [13]. Because of fewer human and financial resources, the introduction of a knowledge management system should not cause ongoing costs. Especially in SMEs there are no specialists for knowledge management and additional staff costs are not manageable. The goal of our project is to account for these special requirements.

In summary, we need to consider the following requirements:

- the effort to record analogue documents has to be minimal and should not disrupt the daily work
- intuitive usability
- there should be no running costs, in particular no additional staff cost
- knowledge should be extracted from different resources
- access to the knowledge base should be possible at any time
- the extracted knowledge should be linked in an effective manner to the usual business processes

During the daily work, the employee can decide whether certain resources should be transferred into the knowledge base. Those resources (e.g. an e-mail or a PDF document) are then passed to the OnToBau-System using the interface of his personal agent or plug-ins integrated in his office or e-mail software.

To transfer analogue documents to the knowledge base, we are planning to install special document cameras directly at the workplace of every employee. In this way we prevent the scanning process from disturbing the workflows in the company and ensure that every employee can easily add new resources to the knowledge base.

The purpose of this section is to give an overview of the architecture of our approach. As shown in Figure 1, the OnToBau-system consists of four main parts:

- preprocessing components,
- information extraction,
- knowledge base,
- a personal agent and user interface.
We describe the architecture and its main features in the following subsections.

**Pre-processing**

The pre-processing components will prepare the resources for inclusion in the knowledge base. Therefore the resources are converted to a general representation language. The OnToBau Representation Language and the general design of the converters will be discussed in more detail in Section 4. The pre-processing also includes various filters (e.g. segmentation filter, part-of-speech tagger, stop word filter), which are used to simplify the subsequent processing. The relevant information is extracted in the extraction system component.

**Information Extraction**

First, a given resource is classified into process related categories (e.g. an invoice). The domain ontologies contain information which relevant data have to be extracted from an invoice. Once the information has been extracted, it is stored in the knowledge base. For this purpose, we try to find relevant relationships between the new data and the existing knowledge in the knowledge base (e.g. a link between an invoice and a corresponding quotation).

**Knowledge Base**

The knowledge in the OnToBau-system is represented by using ontologies. For this, we decided to use the Web Ontology Language (OWL), which is an W3C recommendation since February 2004, so we can benefit from existing libraries and tools. The T-Box of our ontology defines the most relevant concepts and relations of the construction domain and in general of business documents, thus providing the OnToBau-system with the necessary knowledge to decide which information to extract from the resources (e.g. recipient details, product information etc.). The extracted information is stored as individuals in the ontology.

**Personal agent and user interfaces**

The personal agent of the employee performs two main tasks. First, he should allow the employee to access the knowledge base by making specific search queries (e.g. show me all invoices of Mr. Brown with an invoice date later than December 2009). For this purpose we implement a graphical tool for query construction. Second, the agent should monitor the employee’s behavior and provide him pro-actively with information to fulfill his task. By this means the user will be provided with useful information based on the actual context and the current step in the workflow.

**4. OnToBau Representation Language**

In this section the preprocessing components are presented. Their objective is to prepare the different business documents for the information extraction process. To allow a general processing of different resources within a company (E-Mails, calendar entries, papers etc.), an XML-based representation language (ORL) was defined, which is discussed in the last subsection. The following subsection presents the concept of the OnToBau converter, whose purpose it is to transfer different resources into the mentioned ORL.

**OnToBau Document Converter**

The document converter’s task is to transfer the various documents into the OnToBau Representation Language, as shown in Figure 2. The document converter can be developed in different programming languages, but they should have certain features. Therefore, an interface is defined which is to be implemented by all converters. For this purpose it is convenient to use creation of design patterns to develop the document converters. The converter can either be directly included as an extension in the employee’s office and email applications (e.g. Thunderbird or Microsoft Office) or can be made accessible via the personal agent. Therefore, the employee has to drag-and-drop the resource into the personal agent’s UI.

![Figure 2: Converting different resources into the ORL](image)

**OnToBau Representation Language**

According to [12] knowledge within a company may be available in two forms: tacit or explicit. Tacit knowledge is based on the experience, expertise, know-how etc. of a person. Often it is very difficult to represent this form of knowledge formally and hence, it is not subject of this research project. Explicit knowledge in contrast, is in a documented form (e.g. in documents, databases etc.) and is directly accessible to all. The focus in the
OnToBau research project is to extract relevant knowledge out of this documented explicit knowledge within a company. Although explicit knowledge is already described in a formal way, there are still many different options of representation. Information can be present in paper-based documents, electronic text files (like WebPages or PDF documents) or in emails, but also in the address book or the calendar of an employee. This diversity of possible information sources makes it difficult to process them in a general way, because each one has its own characteristics. For this reason, the development of an own general representation language for the different resources seems to be appropriate. We use the Extended Markup Language (XML) to describe the ORL. XML has the advantage of being platform-independent and there are many standardized APIs for XML processing in different programming languages. There already exist several XML-based representation languages for different resources, like the Open Document Format for Office Applications (ODF) for text documents or xCal as an XML-compliant representation of the iCalendar standard. But it is hardly possible to use these languages exclusively in our project. One reason is, that these formats are primarily developed as platform-independent and free exchange format (e.g. for office documents) and do not contain specific information relevant for information extraction. Because we consider also paper-based documents we need information about the layout of the form and the coordinates of words, sentences or paragraphs. By defining our own XML-Schema, we can create the markup language in terms of our specific problem set, rather than relying on a generic set of XML-tags suitable for everyone’s need. Nevertheless, we try to incorporate ideas or parts from the existing standards.

The ORL is divided into different abstraction layers, as shown in Figure 3. Each layer represents information about the resources with an increasing level of detail.

**Figure 3: Overview of the ORL abstraction layers**

**Top Level Layer**: The information unit (IU) is the only element in the top level layer and the root of the ORL. It represents the highest abstraction level of a resource and serves as a container for the elements from the layers below. The IU stores metadata about the resources. This includes the Dublin Core metadata and additional metadata, necessary for our approach (e.g. who entered the resource in the knowledge archive, the place of the physical original resource etc.).

**Intermediate Layer**: This layer divides the resources into four main categories we are planning to support in the

**OnToBau-System**: Media, address book, calendar and those resources containing primarily textual information.

**Document Layer**: This layer specifically describes the various forms of documents in each category. It has the highest level of detail, thus containing information which only occurs in those resources (e.g. the subject of an email, the ID3-Tags of a music file etc.).

The ORL is divided into different namespaces in accordance to the defined resource categories in the intermediate layer. For our approach the textual information in the resources is most important, as it mainly contains the knowledge to be extracted in the information extraction part. For this reason we developed a general text representation language (TRL), similar to the ones used in standardized formats like ODF; providing elements to segment the text in paragraphs, headings etc. and to mark text parts with typographic information like bold text. As shown in Figure 4, an IU consists of three sub-elements: The first one representing general metadata about the resource. The second element representing the textual information extracted from the resource, as original text and in the TRL. At last the third element contains document specific information about the resource.

The different converters produce the corresponding ORL file from a given resource (e.g. a Microsoft Office Word document). This ORL file is the input for the information extraction component in the OnToBau architecture. The advantage of this approach is that the information extraction components can mainly be developed independently of the given resource. This offers the opportunity to easily expand the system with new resources (e.g. Open Office Documents, scanned documents etc.).

![XML representation of the ORL](image-url)

**Figure 4: Example of an ORL-file**
5. CONCLUSION AND FUTURE WORK

Even if the technical results will be applicable for other domains which have quotation preparation processes we focused on the construction domain. Our project partner operates in this domain and supports us with the knowledge to build up the basic domain ontology. Because there are no specific projects supporting the knowledge management process of companies in the targeted domain the presented architecture is our way to provide small and medium companies with an ontology-based knowledge management system. To reach this goal we have developed the ORL. This representation language is the link between different resources used within companies and the information extraction. With specific converters the system is able to transform the resources to a unitary format which is needed for extraction. Currently we are working mainly on the extraction step, to enable the system to build up a knowledge base for a company. This base will be represented by using ontologies. A next step will be achieving some test results with our project partner. This will prove the properness of our approach to build up a knowledge-base in the construction sector.

An integral part of the OnToBau system which is not yet implemented will be the personal user agent. The agent will monitor the user’s activities and provide him with the relevant information for the process in real-time. To achieve this, it needs underlying behavior patterns and must try to anticipate the user’s intention.

6. REFERENCES