Collecting of Content Enterprise Knowledge by using Metadata

Boris Plejić and Zoran Skočir*
Ericsson Nikola Tesla, Networks
Krapinska 45, HR-10000 Zagreb
Faculty of Electrical Engineering and Computing, University of Zagreb*
Unska 3, HR-10000 Zagreb*
Phone : 01-365 3979 E-mail : boris.plejic@ericsson.com

Abstract –The fast development of modern technologies and the all-pervasive globalization of the market are generally accompanied by an increase in the number of processes within a company. Besides physical resources and energy, knowledge is being more and more emphasized as the solution to the modernization of economy and a resource permanently needed in the improving of the quality of business processes. This paper is going to describe the processes of reaching the contained knowledge and depict the system used for collecting the contained knowledge, the one saved in private documents, which are, regarding their content, seen as unstructured sources of knowledge, and which could have a key role in the speeding up of the process, preventing a decrease in the quality of business solutions. Simple model is developed, representing benefits from collecting knowledge by using Metadata.

I. INTRODUCTION

We know that nowadays a company's success on business markets depends on the time of the realization of business solutions, as well as the improvement of its competitivness through quality, so the development of solutions used for collecting contained knowledge should be one of the main factors of doing business and the company's success on the market. Running a business becomes more complex and demanding, while the saving of information and the description of the processes become the key factor in doing business. Special attention in the company is paid to creating, transferring and sharing knowledge among employees, which in the long run leads to a change in the relative value of knowledge and prevents fast destruction of knowledge.

The problem of present-day processes of saving knowledge is that 70% of data within a company are unstructured, i.e. unusable in the processes, 10% of data are semi-structured, i.e. can be used only by some users, whereas only 20% of data are structured, available to all the users, clearly defined and saved. This devastating fact shows that 80% of a company's knowledge, at least regarding its content, is not available to the users and thus represents dead capital. Dead capital hinders a faster increase of the company's value as well as its profits, which directly impacts the company's competitiveness.

The value of a business increasingly lies in intangible assets: companies patents, specific software solutions, research programmes, expertise. Managing any of these assets is very difficult, but the hardest ones to deal with are personal employees knowledge or wisdom. It takes time for wisdom to acquire, to absorb and to record.

Employees knowledge capture will be shown through example of gathering useful informations from MS Word documents using Metadata.

Model will show the creation of object-relational database of integrated Metadata of the company which would ensure informations needed for collecting its organisational knowledge. With the help of Metadata the content within a MS Word document will become accessible.

So we will create a model for transforming unstructured data from scattered sources into new company knowledge.

Paper is structured as follows. In Section II we can find definitions of a term Knowledge. Section III describes known tehnologies for collecting the knowledge. Section IV define term Metadata. In Section V we can find description of companies unstructured data. Section VI offers scenarios for collecting knowledge in the future. In Section VII model for collection unstructured data from MS Word documents is presented. Finally, in Section VIII the conclusion and further works are drawn.

II. KNOWLEDGE

Definitions of knowledge vary. R. Gregory Wenig, (1998) stresses this point [7]:

'Currently, there is no consensus on what knowledge is. Over the millennia, the dominant philosophies of each age have added their own definition of knowledge to the list. It is a construct that is not directly observable'.

But some definitions associated with the word 'knowledge' are given. Here is one from Davenport and Prusak [1]:

'Knowledge is a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information. It originates and is applied in the mind of knowers. In organizations, it often becomes embedded not only in documents and repositories but also in organizational routines, processes, practices and norms'.

Knowledge represents the synthesis of all the information, technical skill, research experience and the estimations which have their value and are structured in one place.

Knowledge can be divided according to content into [8]:

- factual knowledge;
- procedural knowledge;
- knowledge needed for evaluation.

Factual knowledge represents data and informations. Procedural knowledge covers techniques in the field of

heuristic processes and algorithms, processes that are using unproven allegations with a purpose of scientific evidence. Knowledge needed for evaluation is important in the processes of making decisions, when managers need to do the estimations of limitations within particular projects, based on which they set the goals and objectives.

Davenport and Prusak classified knowledge by components to tacit knowledge and explicit knowledge[1]:

'Tacit knowledge is personal knowledge in the form of skills, know-how, experience, intuition, insights, feelings and beliefs'.

'Explicit knowledge is knowledge contained in oral or written language intended for consumption or access by others. It is knowledge that has been formulated and formalised, and is typically found in books, documents, manuals, formulae, presentations, lectures, etc.'.

The information found in the company, from which new knowledge is generated can be divided into: (Figure 1)

- Unstructured data - the form of data which are hard to find, unless sorted out within the contained management and marked with metadata;
  - images (.jpg, .gif), content of Web documents, PDF documents, standard documents (like MS Office documents), text documents, e-mail, Media documents (.mp3, .wma or .wmv) or other audio, video and correspondence;
- Structured data - data defined content-wise at least in the title field, easily accessible for further processing;
  - relational databases, structured data files, system/application data and logs that reside in a data store, defined by a catalog (table definitions)/data model accessible via SQL or Object definitions;
- Semi-structured data - data warehouses structure with freeform elements (e.g., e-mails) and has structure and context to specific elements in the header.

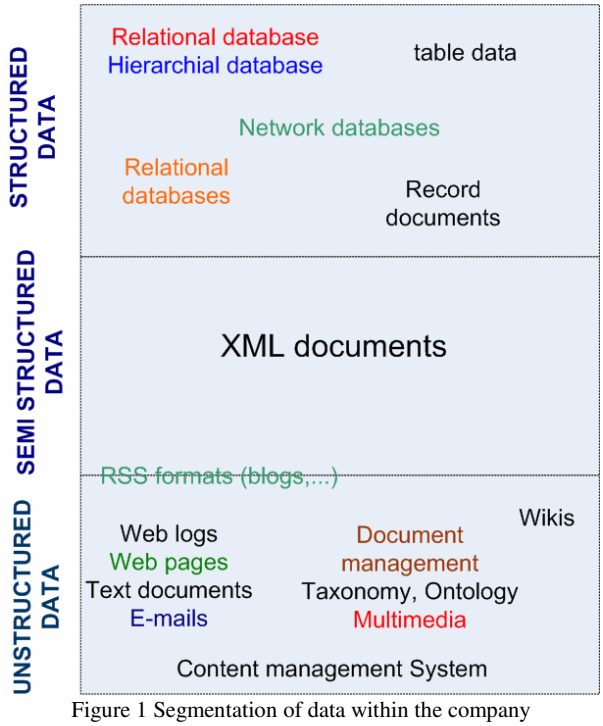

Figure 1 Segmentation of data within the company

In the company data are either saved in the data warehouse (structured data) or scattered throughout departments and among employees ranging from Finances, Human Resources, to Research & Development department (unstructured data). Evaluating knowledge leads to saving important knowledge which appears in the form of checked conclusions and models. In this way repeated use of knowledge leads to greater confidence in the success of the project and more trust in the repeated usability of the knowledge.

According to Nonaka and Konno, knowledge creation is a spiraling process of interactions between explicit and tacit knowledge, described with SECI model (Socialization - Externalization - Combination - Internalization) [11].

Figure 2 shows four conversion patterns of knowledge:

1. Socialization - enables the conversion of tacit knowledge through face-to-face communication;
2. Externalization - requires the expression of tacit knowledge and its translation into concepts that can be understood by others;
3. Combination - involves the conversion of explicit knowledge into various sets of explicit knowledge;
4. Internalization of newly created knowledge is the conversion of explicit knowledge into the organization's tacit assets.

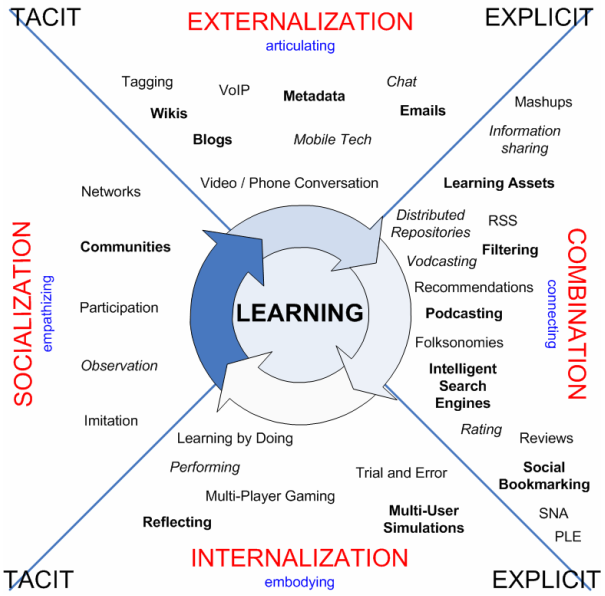

Figure 2 Web 2.0 Driven SECI Model [12]

The procedure called Knowledge Management deals with the evaluating of knowledge. One definition of the term Knowledge Management says [9]:

'Knowledge Management caters to the critical issues of organisation adaption, survival and competence in face of increasingly discontinuous change. Essentially, it embodies organisational processes that seek synergistic combination of data and information processing capacity of information technologies, and the creative and innovative capacity of human beings'.

The system of Knowledge Management within a project is implemented after an overall picture of the particular project is made. In the process the following points have to be paid attention to:

- Short and quick solutions are always scaleable and usable on a higher level, and as such carry a dose of false impression about success. When solving a problem, as much as possible previously acquired knowledge about the process has to be involved;
- Project management office has to represent the base for implementing the system of managing knowledge within a project in accordance with project management;
- To study and analyze the results of the project and follow the achievements and procedures from previous projects;
- The implementation of the solutions should be accelerated and the goal accomplished and analyzed with key indicators of effectivenes called key performance indicators (KPI);
- All the experience acquired in previous projects, presented in factual knowledge, should be available in order to accelerate the process without decreasing the quality.

We can see that it is essential to reach the needed data precisely and fast and in this way achieve a perfect system of managing knowledge within projects, i.e. generate quality knowledge from the data. Nowadays we are succeeding in doing this with only 30% of data within the company (structured and semi-structured data), while other data are mostly unusable for the employees' personal use. (Figure 3)

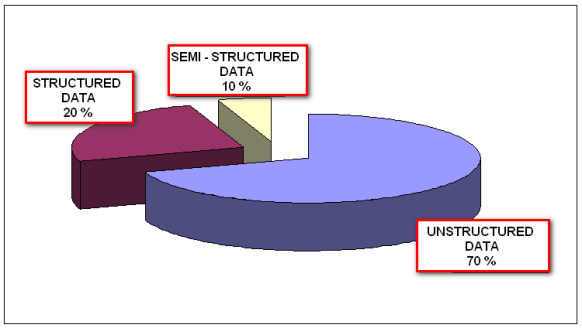

Figure 3 The present structure of the data within the company

III. TEHNOLOGIES FOR REACHING KNOWLEDGE IMPLEMENTED IN COMPANY

For some time organizations have been mostly relying on business intelligence technology (BI) while using structured data, enterprise content management technologies (ECM) while using semi-structured data, unstructured groups of documents, web contents, e-mails and reports.

Definition of BI can be find on Wikipedia, a web-based, free-content encyclopedia:

'Business intelligence refers to skills, knowledge, technologies, applications, quality, risks, security issues and practices used to help a business to acquire a better understanding of market behavior and commercial context. For this purpose it undertakes the collection, integration, analysis, interpretation and presentation of business information. By extension, "business intelligence" may refer to the collected information itself or the explicit knowledge developed from the information.'

Definition of ECM find on Wikipedia is:

'Enterprise content management is the strategies, methods and tools used to capture, manage, store, preserve, and deliver content and documents related to organizational processes. ECM tools and strategies allow the management of an organization's unstructured information, wherever that information exists'.

Integration of this group of values on a portable level has been successfully defined in the past. However, BI technology relies solely on facts mainly saved within special structures, such as relational databases, not the content of documents. Management technologies which use the content of documents, including the tools for reaching, processing, saving and accessing the content of unstructured data are still being developed and have not been completely implemented in the company's processes, which causes companies to have great losses on all levels.

Nowadays in the chaotic business world BI and ECM technologies are not sufficient any more, which is reflected in the more frequent occurrence of mistakes created by the wrong understanding of data, which can be very harmful for the processes. In the processes it is not sufficient any more only to find the saved information, but also use it profitably and at the same time understand the purpose of the data. The purpose of the data is increased with the access to the content of the data and its becoming more available for further usage. Today the rule is that data are saved in the structured form within databases or data warehouse. (Figure 4)

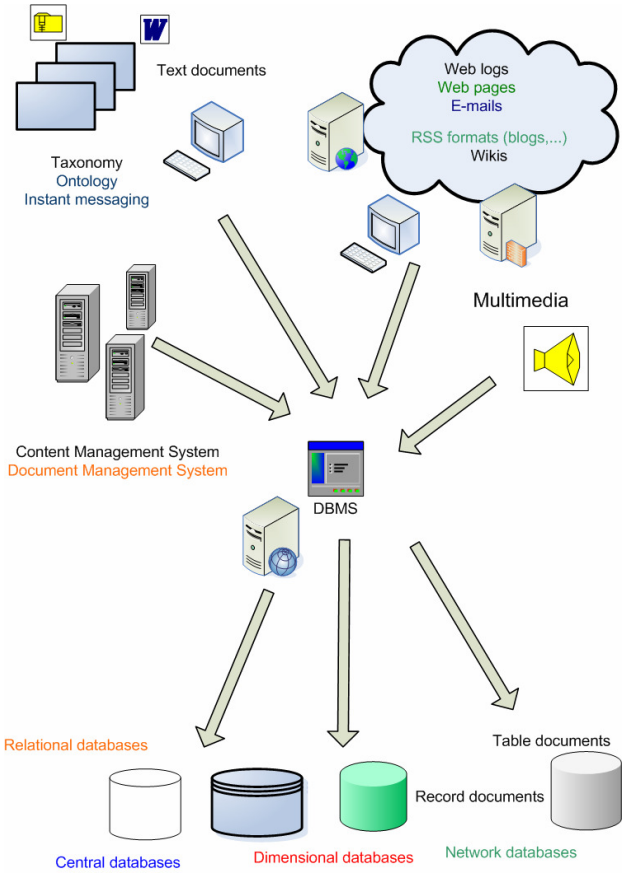

Figure 4 Present data retrieval process from various sources

But if we describe the content of the data (past, business rules related to the data, definitions, characteristics, affiliation) the data becomes clearer. The communication among people about the problem then gets more fluent and the decisions within the process more precise and with a more favorable outcome. The content of the information is then accessible through the description of data, called Metadata, and information which derives from it the company's strategic good.

IV. METADATA

Metadata (data about data) is a term which denotes secondary, auxiliary data which contain data about data or the data as to how to process these data in the easiest way.

Metadata represent the appropriate catalogue about data. The possibility of using knowledge which is accumulated through fast and precise use of data with the usage of Metadata becomes one of the factors of company's making business and it market competitiveness. By enabling the employee to reach the Metadata, new and innovative ways of accessing data are created.

Some of the common places to mine for metadata in company are :

- inside databases:
  - Database catalog, design models and documents;
  - IMS segments, Document Type Definitions (DTDs);
- inside Object-Oriented Applications and Logical Models:
  - Interface Definition Language (IDL);
  - Class definitions;
  - Source code management tools;
  - Entity-Relationship diagrams;
  - CASE tools;
  - Unified Modeling Language (UML) tools;
  - Schemas inside Enterprise Resource Planning Software;
  - Data and object models;
  - MS Office documents (Word, Excel, Access, Visio, Project, ...).

Unstructured data in combination with Metadata become semi-structured data, accessible through index schemes.

Today we often encounter examples of these data within the company.

Thus, a Microsoft Word document is structured by markers, Metadata and process instructions in Word, but its part described by contents is unstructured (mainly the text found in the body of the document), which could contain the description of the process and the significant characteristics of the procedure within the process, or the observations of the employees on previous similar projects. Also, the web page and HTML have structured tags but unstructured text. E-mails are made of structured form through pictures, audio and video data, but the text itself is unstructured.

We can see that each data can be indexed and made accessible through correct formatting, but what to do with the knowledge saved within content-described part of data?

The feeling is as if you are reading only the contents of a book while skipping the text. In this case we have trouble reaching the needed data or we cannot reach them at all. If reading in this way, many things can be understood wrongly.

Processing of the content of unstructured data and its integration are possible only after the content has been pulled out, separated from the data. This is enabled by the integration and definition of Metadata within a document, making the content accessible. We can see that unstructured and semi-structured data are not the same content-wise, but are the same regarding the structural scheme. In this way each unstructured data, and at the same time the information pulled out of the text within data, can become accessible with the help of the Metadata definition and thus 'become' saveable semi-structured data.

And that is the key for collecting unstructured assets and introducing a Metadata into companies knowledge structure.

V. COMPANIES UNSTRUCTURED DATA

Today we have a clear picture of media used for saving data within a company, which mainly consists in relational, hierarchical, multi-dimensional and central databases. Also, here are the data in the form of tables - spreadsheets, network reports and pages, whereas the data saved in the form of XML documents and multimedia documents are represented in lower numbers. Here we are talking mostly about structured or semi-structured data.

When we look towards the future, the majority of researches show a great increase in unstructured data which will need to be saved somehow, which also means a decrease in the amount of structured data saved within data bases. Sound data, blogs, Wikis, network-oriented documents (e-mails), text-oriented documents (MS Office documents) and multimedia documents will encounter the largest rise.

The decrease in structured data over a few years is 15 to 46%, whereas the expected rise in the unstructured data ranges between 61 to 81% [10].

Soon it will be normal to save unstructured data within databases.

One of the main differences between structured and unstructured data is periodic renewal of the environment of structured data (databases and data warehouses), with every change being followed, with data being updated and a trace left in the change. However, with majority of data within a company this is not the case.

Unstructured data are mainly changed only after having been defined. For example, after an e-mail is written and sent, it cannot be changed any more. After a Word document has been written and released, it cannot be altered any more.

Therefore there occur great discrepancies between the environement of the structured data and that of the unstructured data, which will be explained in more detail in Section VI.

Interesting books about the companies unstructured and structured environmetn and usage of Metadata and KM can be found in [2, 3, 4 and 5].

VI. THE FUTURE OF COLLECTING KNOWLEDGE

Collecting unstructured data will in the future require the adjusting of databases and Data Base Managment Systems i.e. DBMS, because there has to occur the reduction of the saveable data format. In this way the number of kinds of databases will be reduced, as well as table data within relational databases, which could cause problems with relational database, where table data are integrated in high degree. There exists a need for implementing a robust system for processing Metadata which would be used for saving data in the data warehouse. (Figure 5)

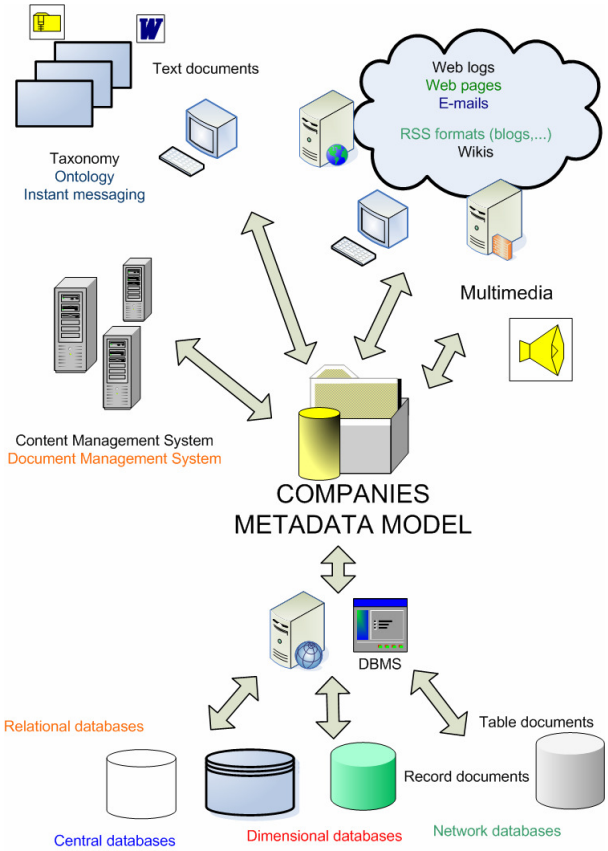

Figure 5 Enterprise Knowledge Management model

Evolution will eventually lead to the transformation of unstructured data in data formats which are readable for data warehouses with the help of Metadata, and as such will be saved into data warehouses, which can also expect changes. Current interfaces for saving structured data within data warehouses will have to adapt to the accessing of unstructured data through a transformation of these data prior to their being saved in the data warehouses. Also, data models within data warehouses will face certain changes and expansion. This adaptation should proceed smoothly since the data coming into the warehouse should already be structured.

Employees working on data warehouses will have to have additional training since there is a new scenario of changing data in the company.

Unstructured data, shown in the format which is saveable in the data warehouse, should then be updated and periodically renewed, and these changes should be followed.

Nowadays the processes of analysis are based on business analysis of structured data because they are marked by a transaction or numerical data and thus naturally connect the present analysis processes.

The processes of analysis will in the future have to solve the problems which will be caused by the rise in unstructured data, such as:

- Unstructured data can be found in many various formats, with some of them being more and some less accessible. The processes will have to reach any data regardless of their accessibility;
- A large problem will be the terminology because it is common for different users to call the same unstructured data differently;
- The size of the data is a potential problem because unstructured data have to be examined in totality and not only by the title, which is the case with structured data. Searching the text of the data with the help of Metadata thus becomes essential;
- Unstructured data will have to be divided according to their priorities, which could be a problem when there are more users of a particular data, and for whom it is more or less important;
- The costs of integrating new solutions within the existing mechanisms of analysis;
- Implemeting of a higher level of safety within the system for the processing of unstructured data.

The identification and cataloguing of unstructured data creates a model which is really an active management system of Metadata and is used by the organization in a vigorous processing of Metadata regardless of many obstacles encountered when scattered Metadata are saved within a single Metadata model. In this model Metadata would describe well the data and be automatically refreshed by the system!

VII. EXAMPLE OF ACCESSING DATA FROM THE TEXT USING METADATA

Data and information comprising knowledge of the company are most often scattered among employees, projects and numerous warehouses where they are saved in various forms, mostly as documents. Companies which strive to effectively use and mobilize the knowledge with the goal of improving the business make an effort to organize the system for using knowledge saved in various documents.

A study example (Figure 6) shows the creation of object-relational data base of integrated Metadata of the company which would ensure relevant information needed for following its organisational knowledge. With the help of Metadata the content within a Word document, which comprises requirements defined by the Project Management, becomes accessible and the knowledge, i.e. information, is used for creating a team competent for fulfilling the requirement.

Within a database we have a table *MetaDataTable* with Metadata accessible from a set of Word documents.

The first macro accesses Metadata and goes with them as with an input through Word documents and pulls out data described by Metadata. The data accessed in the Word

documents it saves in a database within the table *Zahtjevi* and also into a Word document *Prometi.doc* which is used for the further analysis of the transactions and the changes made to the accessed documents.

Then the second macro, according to requirements (ranked according to priority), selects employees which are capable of fulfilling the requirements and saves them in a new table *Rezultat* within the database. In the process the employees can be assigned to a maximum number of projects defined in the variable *maxbrojprojekata*. The value of the variable *maxbrojprojekata* is determined in the table *MetaDataTable*.

The third macro draws out the table from the database in a Word document *Rezultat.doc* which contains the information required by the management. By using the Metadata the access to the needed information is possible regardless of where in the contents the information is situated, if it is only correctly defined in the table *MetaDataTable*. In this way the unstructured data (information) becomes accessible through Metadata.

If the data is not defined by Metadata it is enough to change its description for it to be lost!

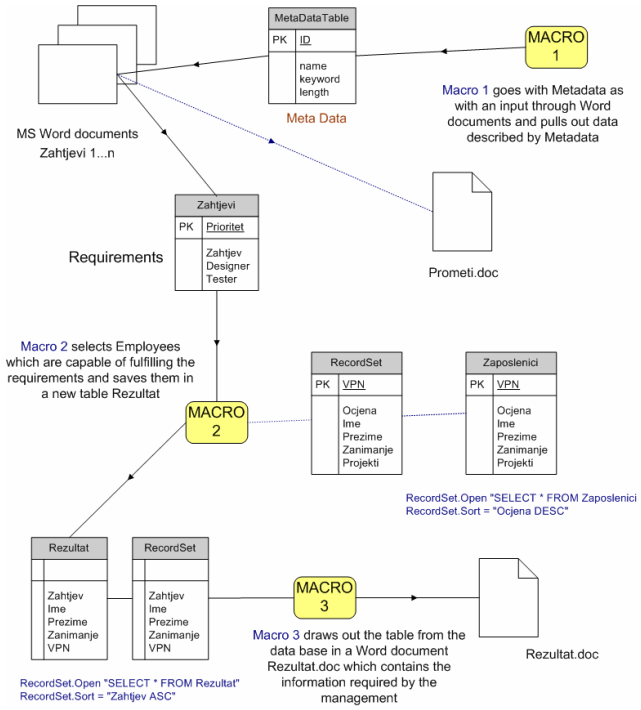

Figure 6 Case study

VIII. CONCLUSION

Nowadays companies have advanced technologies that can increasingly accomplish programmed processes traditionally done by humans, like taking out the informations from scattered unstructured sources. The challenge for companies is to find better ways to extract and find such valuable information.

English author, courtier, and philosopher Francis Bacon said that ‘Knowledge is power’ [6].

With help of New tehnologies 'power of knowledge' is growing and lost companies assets are stored and set for creating usable knowledge management and a new companies wisdom. New technologies must reinforce existing human rules of knowledge management without replacing employees force. Technology is part of the answer, and managerial ingenuity must do the rest.

In this paper we propose use of Metadata benefits to business needs. Using and developing simple tools for collecting unstructured data companies can relatively easy create new wisdom and reduce companies scattered unused assets. Our main obligation must be collecting of unstructured data (companies dead capital) and creating a new knowledge values to gain a respectable advantage on the market.

Further works can concentrate on building architecture where all unstructured data will be fetched and metamorphosed to structured assets.

REFERENCES

[1] T. H. Davenport and L. Prusak, *Working Knowledge: How Organizations Manage What They Know*, Harvard Business School Press, Boston, 2000.

[2] W. H. Inmon, B. O'Neil and L. Fryman, *Business Metadata: Capturing Enterprise Knowledge*, Morgan Kaufmann Publishers, Burlington, 2007.

[3] K. Hammer and T. Timmerman, *Fundamentals of Software Integration*, Jones and Bartlett Publishers, Sudbury, 2007.

[4] P. Gottschalk, *Knowledge Management Systems: Value Shop Creation*, Idea Group Publishing, Hershey, 2007.

[5] W. H. Inmon and A. Nesavich, *Tapping into Unstructured Data: Integrating Unstructured Data and Textual Analytics into Business Intelligence*, Prentice Hall, Boston, 2008.

[6] B. Francis, *Essayes. Religious Meditations. Places of perswasion and disswasion*, printed for H. Hooper, London, 1597.

[7] J. M. Firestone, *Enterprise Information Portals and Knowledge Management*, Butterworth-Heinemann, Burlington, 2003.

[8] S. Vidović, “Upravljanje znanjem”, InfoTrend online num. 107, Info Press, Zagreb, 2003; http://www.trend.hr/clanak.aspx?BrojID=5&KatID=5&ClanakID=139.

[9] R. Rao, "From Unstructured Data to Actionable Intelligence", *IT Professional*, p. 29-35, 2003.

[10] P. Russom, "TDMI Research : BI Search and Text Analytics", *DM Review Special Report*, 2007.

[11] I. Nonaka and N. Konno, "The concept of "Ba': Building foundation for Knowledge Creation.", *California Management Review Vol 40*, 1998.

[12] M. A. Chatti, R. Klamma, M. Jarke, A. Naeve, "The Web 2.0 Driven SECI Model Based Learning Process", *Proceedings of the 7th IEEE International Conference on Advanced Learning Technologies, ICALT 2007*, IEEE Computer Society, July 2007.