

**mag. Stojan Košti, doc. dr. Bojan Cestnik,**

**Temida d.o.o.**

**Dunajska 51, 1000 Ljubljana, Slovenia**

**Tel.: 012363350, Fax.: 012363356, e-mail: stojan.kosti@temida.si**

---

## **Structuring Domain Knowledge by Ontology Construction**

***Abstract:** In the paper we first give an overview of the literature mining technology. Then we list its main application areas with a special emphasis on the role of learning ontologies from texts. In order to understand how organization functions to accomplish its purposes we use some Business Analysis approaches to. Representation of learned concepts in the form of ontology is a basis for structuring domain knowledge and facilitates better communication with domain experts. To pursue this mission, analysts have to collect, analyze and synthesize huge amount of information. Ontologies are suitable also for effective apprehension of a target problem domain, since they comprise generalized view on the studied literature. For the case study we take a set of abstracts from the articles presented on CompSysTech'09 conference and use OntoGen tool to semi-automatically construct an ontology for the domain. The obtained results indicate that the proposed approach can be effectively used to digest textual information and present it in a more operational form of topic ontology.*

***Key words:** ontology, information, data, text mining, domain knowledge*

## **Semantično iskanje znanja v dokumentih s pomočjo gradnje ontologij**

***Izveček:** V članku najprej na kratko predstavimo tehnologijo iskanja znanja v besedilih. Nato podajamo in opišemo glavna področja uporabe, pri čemer posebej izpostavimo vlogo učenja ontologij iz dokumentov. Z namenom širšega razumevanja in umeščanja ontologij z vidika funkcioniranja organizacije se naslanjamo tudi na metode poslovne analitike. Za doseg cilja je namreč potrebno zbrati, analizirati, urediti in semantično organizirati veliko količino informacij. Predstavitev naučenih konceptov v obliki ontologije je osnova za strukturiranje znanja na izbranem problemskem področju in omogoča boljše sporazumevanje s strokovnjaki s posameznega področja. Ontologije so primerne tudi za učinkovito seznanjanje s problemskim področjem, saj vsebujejo splošen pogled na obravnavano množico besedil. Uporabnost tehnologije rudarjenja v besedilih podrobneje ilustriramo na dveh primerih: pri odkrivanju novih povezav v literaturi in pri izboljšanju razumljivosti besedil.*

***Ključne besede:** ontologija, informacije, podatki, iskanje znanja, poslovno področje*

## **INTRODUCTION**

The matter of ontologies belongs to the branch of philosophy that deals with theories about the structure and behavior of the worlds that humans perceive. Strictly speaking, ontology is a philosophical science concerned with what exists: the science of “being”. For the purpose of this paper, however, ontologies are used to model the knowledge in a particular domain. They include domain concepts we can discuss, their properties and relations with each other. Ontologists seek to articulate the fundamental types of phenomena that exist in the world and the relationships that can arise among these different types of phenomena [1].

The quantity of knowledge stored in the form of scientific publications, articles and books reveals an exponential growth in the recent years. Consequently, the interest in using text-mining technologies evidently rises, primarily because effectively following the progress in large quantity of published knowledge is almost impossible even in a relatively narrow field of interest. Written text typically contains large quantity of information that is usually encoded in a way hard to comprehend for automatic approaches. In spite of that fact, information technology spawned, besides the area of text mining, several important application areas like information retrieval, computational linguistics, text categorization, ontology learning and hypothesis formation. In the following paragraphs each of the enumerated areas are shortly described and compared to text mining.

Business Analysis (BA), on the other hand, is a modern discipline that is involved in most projects where efficiency is to be improved. The tendency of BA is to improve a value for a customer, sometimes even acting as change agents, which in turn brings more revenue to an organization [2]. Typical sources for finding suitable information are scientific and business conferences addressing the similar problem. Conference papers and presentations offer a comprehensive picture of the involved area; addressed problems are usually selected from the front edge of the current state of affairs. The dilemma that we all face in the modern era is how much time shall we devote to reading all the relevant papers of interest in the Conference Proceedings.

Any tool that might improve our efficiency in this matter is of great value to a typical modern researcher and analyst. In this paper we address the following problem: can we use existing semi-automatic ontology construction tools to review substantial amount of material in the form of abstracts to find relevant information? The information retrieval pursues a goal of helping to find documents that correspond to user’s search requests and criteria. Its main application area is document searching over the Internet. The main difference between text mining and information retrieval lies in the fact that the text mining endorses discovery of new knowledge and searches for typical patterns within collections of text documents, while information retrieval only effectively searches text documents to identify those that meet user’s search criteria. The later, therefore, does not imply searching for new knowledge; it merely focuses on displaying the information that is already present in a collection of documents, but is difficult to find because of the sheer size of the collection. In fact, if we are going to conduct a usability study, it is advisable to find out whether there is any new research going on about the activity of interest [3].

As a case study as well as an input set of documents, we took abstracts from CompSysTech’09 conference proceedings [7]. Suppose that we first want to get an overview of the addressed topics. Then, we want to find and read only the papers that address issues of interest and deal with problems that are related to the domain area under our consideration. As business analyst we should examine, collect and analyze information from various sources, examine similar cases of substantiated findings. We need to establish appropriate proposals that comply with corporate business strategy and business model. On this basis we propose an effective solution for all stakeholders and participating partners.

The main purpose of this article is to provide an overview of the usefulness of tools to assist in the construction of semantic ontologies. First, we describe the use of ontologies as tools for modeling domains of interest. Then, we describe the experimental setup for semi-automatic ontology construction and present the results of our experiments. Next, we demonstrate how ontology construction and text mining approach can be used to improve comprehension of written documents. Finally, we conclude by pointing out the most important findings in the paper.

## **ONTOLOGIES AS DOMAIN MODELLING TOOLS**

Ontologies, explicit formal specifications of terms and relationships among them, play a fundamental role as enabling method for the description of semantically rich, machine-processable business process definitions. Traditionally, ontology construction is a manual task that consists of determining interesting domain concepts and establishing a hierarchy of such concepts. The process uses a special sort of description language, in which common domain knowledge is then represented. Representing business process definitions and related artifacts (e.g. business documents, business objects, etc.) as knowledge, based on ontologies, is a novel approach. Since manual ontology construction is a complex and intricate process that requires both skill and diligence, the need for developing more active and helpful computerized support is evident. To remedy the situation, business and IT must be able to speak the same language and share a common understanding of the business vocabulary and grammar. Business analysts will take on the challenging task of defining the semantics.

With the emergence of new text mining technologies ontologies can be constructed semi-automatically by processing available text documents. In the last decade several approaches for facilitating semi-automatic construction of ontologies have been developed and successfully used in practice, making the process of ontology construction more effective and viable. One example of a tool for interactive construction of ontologies from text documents is OntoGen [3], which has already been proven successful in several real-world applications. A user can form concepts, edit them thematically and assign documents to the formed concepts. By implementing several modern machine learning techniques OntoGen helps users in all crucial phases of ontology construction, suggesting concepts and their names and automatically assigning documents to the proposed concepts (Fortuna, 2006).

The process of ontology construction from a set of documents can be further automatized by using new text mining technologies. In this context, ontologies can be constructed semi-automatically by processing available text documents. Manual ontology construction is a complex and intricate process that requires both skill and diligence. Therefore, there is evident need to develop more active and helpful computerized support for the task. One of the approaches for facilitating semi-automatic construction of ontologies from text documents resulted in a tool OntoGen [6]. Using OntoGen a user can form concepts, edit them thematically and assign documents to the formed concepts. By implementing modern machine learning techniques OntoGen can help users in all crucial phases of ontology construction, suggesting concepts and their names and automatically assigning documents to the proposed concepts [6].

## **EXPERIMENTAL SETUP**

In the domains of interest, knowing domain knowledge structure is beneficial for further learning and analysis. Typically, we start by browsing a set of documents that describe the target domain. The task is to grasp the relevant concepts that are meaningful for structuring the domain knowledge to obtain an overview. Then, depending on our further interest, we usually dig deeper by thoroughly reading a few selected documents that match certain criteria. So, we partially specialize our knowledge in the areas that we find interesting.

Note that the first step (overview) is necessary to improve the effectiveness of the second step (specialization).

To illustrate the procedure, we took 93 titles and abstracts of the papers published in the CompSysTech'09 proceedings as a set of input documents. Our motivation was construct a top-level ontology by using OntoGen. In the conference proceedings, each paper belongs to a particular clearly marked section. The sections correspond to original topics of the conference CompSysTech'09 and are show in Table 1. For our experiment, we decided to disregard the original conference topics and use OntoGen to construct ontology concepts based on the similarity of the 93 papers.

Paper topics (sessions in the proceedings)	No. of papers
Computer systems (Hardware and software)	7
Computer technologies	13
Application aspects of computer systems and technologies	44
Educational aspects of computer systems and technologies	17
Biometrics	3
PhD Workshop	9

Table 1: Paper topics from CompSysTech'09 conference with the number of papers.

As a first insight in the domain under investigation, OntoGen offers a two-dimensional map of the input documents shown in Figure 1. The numbers in the figure denote documents, while displayed words stand as keywords that best describe that area in the map. By selecting a group of documents from the map as shown by shaded ellipse in the figure, we can see a list of keywords in the highlighted rectangle that best describe the covered documents. Note that in the Figure 1 the shaded area covers the papers that are related to learning.

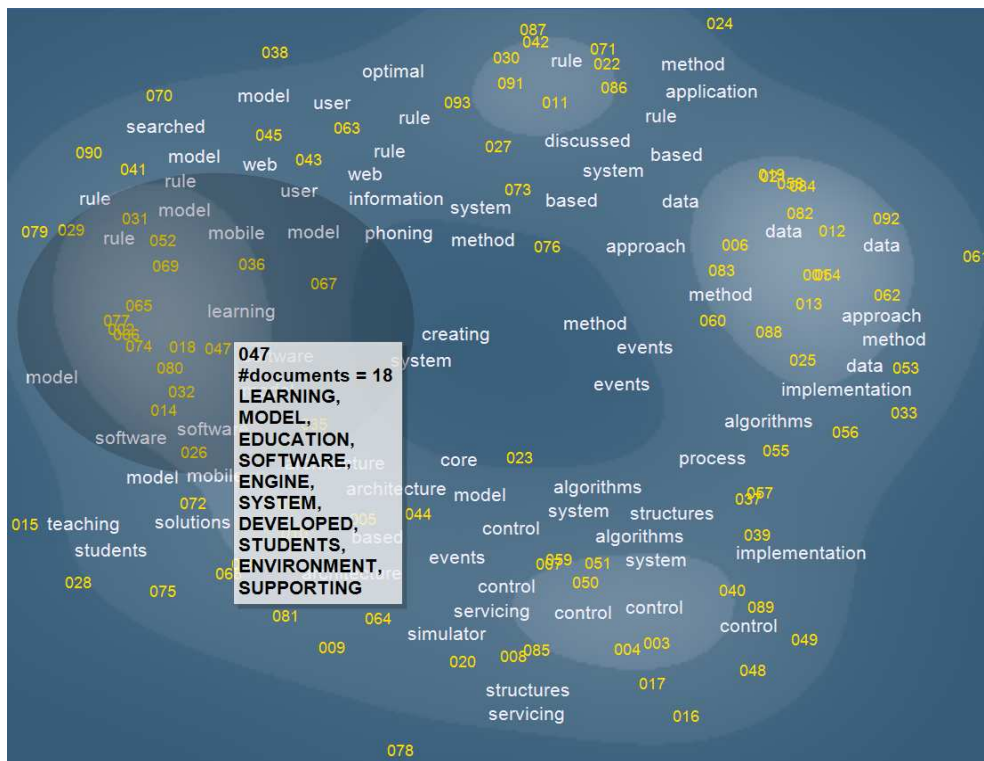


Figure 1: Document map showing the clusters of the input documents and connecting keywords. In the highlighted rectangle, keywords describing documents from the selected area are displayed.

## ONTOLOGY CONSTRUCTION AND EVALUATION OF RESULTS

Top-level topic ontology constructed from the titles and abstracts of the papers published in CompSysTech'09 proceedings [7] is shown in Figure 2. The main concepts are described with three most descriptive keywords. Sometimes it is beneficial if we substitute the three-word description with equivalent phrase that is more commonly understood. This step includes high degree of user's creativity. For example, the concept described by (problems, learning, optimal) can be substituted by the phrase "educational aspects". Such phrase relates the constructed topic more closely to topics from Table 1, enabling easier comprehension and evaluation of the constructed ontology concepts.

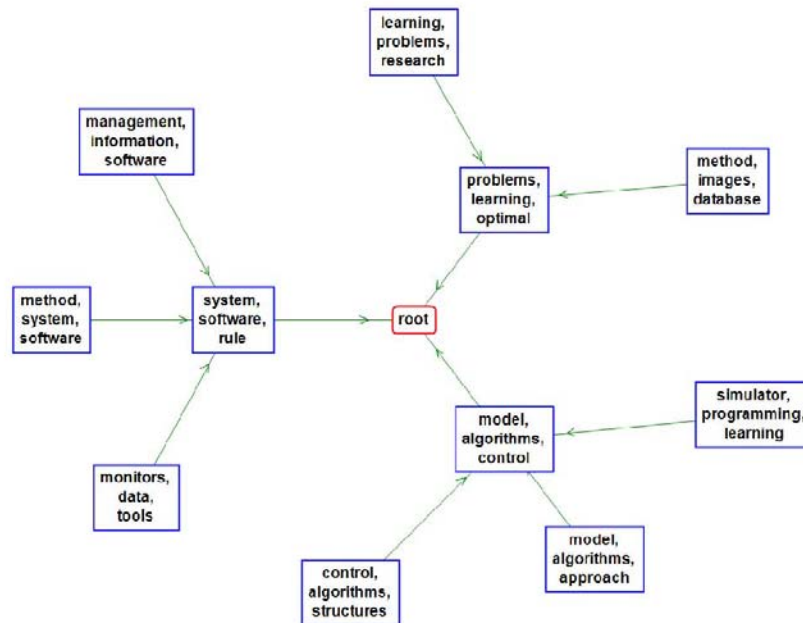


Figure 2: Topic ontology of concepts.

Descriptions of concepts in Figure 2 offer a different view on the conference papers than the topics from Table 1. This is not surprising, since the papers are first compared and clustered syntactically and not semantically. Semantic meaning is assigned to the generated clusters by labeling them with descriptive keywords or concept names. Such concepts offer a "birds-eye" view on the domain, providing a structure of folders for the input set of papers. In such way it enriches our prior knowledge about the domain, motivating creative thinking and additional explanations of the constructed concepts. Therefore, it can be viewed as a tool to enhance learning in the domain under investigation.

When comparing papers that belong to a particular ontology topic, a similarity graph shown in Figure 3 might come in handy. Suppose that the scope of our investigation is the concept labeled (problems, learning, optimal). The papers that belong to its sub-concept (method, images, database) are shown in the central part of the Figure 3 labeled with a checkmark. The same papers appear as red dots in the similarity graph in the lower-right portion of the Figure 3. Note that in this graph the documents are ordered according to their cosine similarity [6]. The thickened blue dot belongs to document 018 [8] that does not belong to the selected concept; it belongs to sub-concept labeled (learning, problems, research). However, it is most similar to the other documents from the concept and can, therefore, be treated as a boundary paper (paper on the edge of a concept). It turned out that studying such boundary papers can be beneficial for shaping the definition of the underlying domain concepts and sub-concepts, which can be useful also for BA activities.

For the purpose of BA, such a method for ontology construction can substantially speed-up the learning process in a given domain. Business analyst can obtain a quick insight in the structure of the target documents. As a result, she can devote more time to identifying and studying interesting sub-topics in more detail.

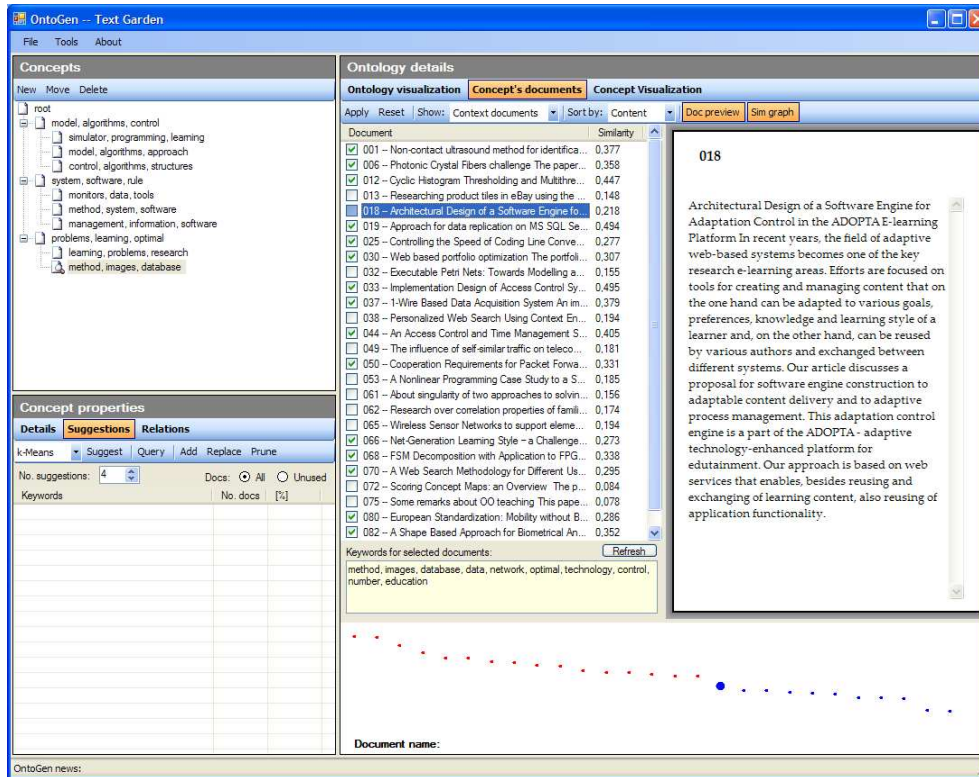


Figure 3: Document similarity graph as depicted by OntoGen.

## SUMMARY

Ontologies provide effective and efficient means for the definition of semantically rich business processes and related entity types. One of the striking benefits lies in the fact that ontology can be fully tailored to the needs of the organization. We demonstrated how such technology could help us answering questions like, for example, which are the areas of research in a given field, in which areas the research activity is the most intensive, how could the domain knowledge be structured into ontology. As a result, we can significantly speed-up the process of “getting acquainted” with a given problem domain, mostly because the generated birds-eye view quickly reveals the most important top-level domain concepts. Ontologies and knowledge bases can be created and maintained by non-IT people, provided that they have received adequate training.

The proposed approach of semi-automatic ontology construction depends on user’s subjective view on the investigation domain. Ontologies built with OntoGen, as an example shown in Figure 2, are constructed in such way that topic suggestions have to be approved by the user. On one hand, the involvement of the user can be considered beneficial since it is actively involved in the creation of the results. On the other hand, however, we have to be aware that the judgments of the user can be subjective in nature. Another user might have perhaps acted differently in a given situation. This fact has to be taken into account when evaluating and acting upon the obtained results.

There is, of course, much more work to be done to develop and test the ideas described in the paper. For example, better tools for process analysis and editing need to be created, more information content needs to be added and systematic tests of how the ideas can be applied in different kinds of situations need to be performed. A business analyst would instantiate ontology, thus creating a knowledge base, which contains comprehensive definitions of business processes and related artifacts. Various artifacts can be generated from the knowledge base, which can also serve as the source for the generation of process definitions. The business analyst is primarily concerned with defining entities that are related to business processes: business activities, business objects, business documents, business rules and roles. Over time, as the number of entities continues to grow, the business analyst will be able to re-use existing entity definitions. However, we believe that our work so far has demonstrated the basic feasibility and contribution of the approach and its potential for significant further progress. We hope, for example, that this research will provide a set of intellectual tools and an extensive database to help people learn about organizations, invent new kinds of organizations, and improve existing processes. Perhaps most importantly, we hope this research will help us understand the possibilities for creating new kinds of processes that are not only more effective but also more fulfilling for their beneficiaries.

## **POVZETEK**

Ontologije so učinkovito sredstvo za opredeljevanje semantično bogatih poslovnih procesov in z njimi povezanih entitet. Ena od večjih prednosti je v tem, da so lahko ontologije v celoti prilagojene potrebam organizacije. V prispevku smo dokazali, kako nam lahko takšna tehnologija pomaga pri odgovorih na vprašanja, ki se pojavljajo, med drugim tudi na področjih raziskav nekega poslovnega področja, kjer je raziskovalna dejavnost najbolj intenzivna in je potrebno znanje ustrezno strukturirati. Rezultat tega je bistveno pospešen proces pridobivanja znanja s posameznih problemskih področij. Predvsem zaradi omogočenega pogleda od zgoraj, lahko hitro razkrijemo najpomembnejše vzorce poslovnega področja na vrhnjem nivoju. Take ontologije in osnove znanja je možno pripraviti in vzdrževati brez poglobljenega znanja informatike, seveda pod pogojem ustrezne usposobljenosti.

Predlagani pristop polavtomatske gradnje ontologij je odvisen od subjektivnega pogleda uporabnikov na raziskovanje poslovnega področja. Pri ontologijah zgrajenih s pomočjo orodja OntoGen, kot je na primer prikazano na sliki 2, mora temo iskanja in raziskovanja predlagati uporabnik sam. Po eni strani je sodelovanje uporabnika koristno, saj je aktivno vključen v oblikovanje rezultatov. Po drugi strani pa se moramo zavedati tudi možnosti subjektivnih presoj uporabnika. V neki dani situaciji bi lahko nek drug uporabnik morda ravnal drugače. Pri ocenjevanju na ta način pridobljenih rezultatov je potrebno to dejstvo tudi upoštevati .

Seveda bo potrebno za razvoj in preizkušanje takih idej opraviti še veliko dela. Potrebno je na primer ustvariti boljše orodja za procesno analizo in urejanje vsebin, v luči kakovostnejših rezultatov je potrebno obdelati več vsebin informacij ter sistematičnih testov o tem, kako se lahko zamisli uporablja v različnih situacijah. Poslovni analitik lahko ontologijo uporablja za ustvarjanje baze znanja, ki vsebuje celovite opredelitve poslovnih procesov in s tem povezanih artefaktov. Poslovni analitik se namreč v prvi vrsti ukvarja z določanjem zadev, ki so povezane s poslovnimi procesi: poslovne dejavnosti, poslovni objekti, cilji, poslovno dokumentacijo, poslovna pravila in vloge. Ker v daljšem časovnem obdobju ponavadi število takih zadev narašča, se lahko obstoje definicije zadev ponovno in večkratno uporabi. Prepričani smo, da je naše delo do sedaj pokazalo osnovno izvedljivost gradnje semantičnih ontologij iz dokumentov in kot tako prispevalo k potencialu prikazanega pristopa za njegov nadaljnji napredek. Upamo, da bodo take in podobne raziskave zagotovile večji

nabor orodij in obsežnejšo bazo podatkov, z namenom lažjega in hitrejšega spoznavanja problemskih področij organizacije, iskanju novih in izboljšanju obstoječih procesov. Morda še pomembneje, upamo, da nam bodo podobne raziskave pomagale razumeti možnosti za oblikovanje novih procesov, ki so ne samo bolj učinkoviti, temveč tudi bolje izpolnjujejo dane zahteve.

## **REFERENCES**

- [1] Green, P., Rosemann. M. Business Systems Analysis with Ontologies. IGI Global, 2005.
- [2] International Institute of Business Analytics. A Guide to the Business Analysis Body of Knowledge Version 2.0. IIBA Toronto Canada, 2009.
- [3] Courage, C., Baxter, K. A Practical Guide to User requirements Methods, Tools, and Techniques. Elsevier Science and Technology Books Inc., 2005.
- [4] Jenz, E.D. Business Process Ontologies: Speeding up Business Process Implementation, Jenz & Partner Gmbh. Erlensee, 2003.
- [5] Gennari J., M.A. Musen, R.W. Ferguson, W.E. Grosso, M. Crubezy, H. Eriksson, N.F. Noy, S.W. Tu. The Evolution of Protégé: An Environment for Knowledge-Based Systems Development. 2002.
- [6] Fortuna B., M. Grobelnik M., D. Mladenić. System for semi-automatic ontology construction. Demo at ESWC 2006. Budva, Črna Gora, June, 2006.
- [7] Rachev B., A. Smrikarov. Proceedings of the International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing 2009, Ruse, Bulgaria, June 18 -19, 2009.
- [8] Bontchev B., D. Vassileva, B. Chavkova, V. Mitev. Architectural design of a software engine for adaptation control in the ADOPTA e-learning platform. Proceedings of the International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing, Ruse, Bulgaria, June 18 -19, 2009.

## **About the author**

**Doc. dr. Bojan Cestnik** is the general manager of software company Temida and a researcher in the Intelligent Data Analysis and Computational Linguistics Research Group at Jožef Stefan Institute in Ljubljana. He is also a professor at the University of Nova Gorica. He obtained his Ph.D. in Computer Science in 1991 at the Faculty of Electrical Engineering and Computer Science, University of Ljubljana, Slovenia. His professional and research interests include knowledge based information systems, decision support systems and machine learning. His research work was presented at several international conferences. He has been responsible for several large-scale software development and maintenance projects.

**Mag. Stojan Košti** is expert covering many levels of analyzing and designing information systems. As an information architect he is involved in particular stages of analyze and design seeking to find out proper Enterprise Architecture to improve performance, security and robustness of IT networks, systems and applications. As a business and system analyst, he accomplished in full life cycles on numerous projects from initial conception to completion from customer liaison, process and systems analysis, design and development, programming, testing to implementation. His research is focused on field of business analysis. He is an author of several professional articles and lecturer on national and international conferences.

## **O avtorjih**

**Dr. Bojan Cestnik** je direktor podjetja Temida d.o.o. Je izkušen sistemski analitik in programer zahtevnih računalniških aplikacij z dvajsetletnimi izkušnjami. Dopolnilno je zaposlen kot raziskovalec na Institutu Jožef Stefan. Je tudi predavatelj na Univerzi v Novi Gorici. Njegovo strokovno in raziskovalno delo so na znanju temelječe informacijski sistemi, sistemi za podporo odločanju in strojno učenje. Njegovo raziskovalno delo je bilo predstavljeno na številnih mednarodnih konferencah. Odgovoren je bil za več obsežnih projektov razvoja implementacije in vzdrževanja programske opreme.

**Mag. Stojan Košti** je izkušen poslovni in sistemski analitik. Njegovo delo je usmerjeno v načrtovanje in oblikovanje spletnih servisov in portalov, ter informatizacijo poslovnih procesov s poudarkom na inovativni uporabi spletnih tehnologij in tehnologij e-poslovanja. Kot informacijski arhitekt je bil vključen v številne projekte v fazah analize in načrtovanja z namenom izboljšanja in optimiziranja organizacijskih in informacijskih rešitev. Raziskovalno se ukvarja s področjem poslovne analitike. Je avtor več strokovnih prispevkov in predavatelj na domačih in mednarodnih konferencah s strokovno tematiko.