Title	Constructing an Ontology for a Research Program
Author(s)	Tian, Jing; Wierzbicki, Andrzej P.; Ren, Hongtao; Nakamori, Yoshiteru
Citation	
Issue Date	2007-11
Type	Conference Paper
Text version	publisher
URL	http://hdl.handle.net/10119/4136
Rights	
Description	The original publication is available at JAIST Press http://www.jaist.ac.jp/library/jaist-press/index.html, Proceedings of KSS'2007 : The Eighth International Symposium on Knowledge and Systems Sciences : November 5-7, 2007, [Ishikawa High-Tech Conference Center, Nomi, Ishikawa, JAPAN], Organized by: Japan Advanced Institute of Science and Technology

Constructing an Ontology for a Research Program

Jing Tian†, Andrzej P. Wierzbicki †‡, Hongtao Ren*, Yoshiteru Nakamori †*

†Center for Strategic Development of Science and Technology,
Japan Advanced Institute of Science and Technology (JAIST)
‡National Institute of Telecommunications, Poland
*School of Knowledge Science, JAIST
{jtian, andrzej, hongtao, nakamori}@jaist.ac.jp

Abstract

In the last decade, the word "ontology" has become a fashionable word in Knowledge Engineering area. The necessity of an ontology and ontological engineering is well-understood. This paper presents an attempt to create an ontology characterizing the 21st Century COE Program *Technology Creation Based on Knowledge Science* at Japan Advanced Institute of Science and Technology (JAIST), based on a combination of a bottom-up and top-down approaches to ontology creation. An example of application of this ontology, related to an *adaptive hermeneutic agent* (AHA), is given.

Keywords: Ontology, Knowledge Science, Knowledge Engineering, Knowledge Management, Technology Management

1. Introduction

In the original sense, ontology is a term in philosophy and its meaning is "theory of being" or "theory of existence", see, e.g., Heidegger [1]. Ontology, however, is also given diverse other meanings, such as 1) a representational vocabulary which can be specialized to some domain or subject matter, 2) a body of knowledge describing some domain. In the context of knowledge sharing, ontology is an explicit specification of conceptualization. In contemporary computer science, ontology is defined as a formal language-like specification of a domain knowledge – actually equivalent to a taxonomy of concepts in a given field of knowledge, enhanced by a structure of hierarchical dependences and other links between concepts constituting the taxonomy, see, e.g., Dieng and Corby [2]. *Ideally, an ontology should provide:*

1) *a common vocabulary,*
2) *explication of what has been often left implicit,*
3) *systematization of knowledge,*
4) *standardization of terms,*
5) *meta-model functionality* (providing a metalanguage for specific models in the domain).

Actually, these goals are not attainable: in order to have a formal meta-model [3], we need a meta-meta-model and so on, therefore we have to stop at some level of explication of basic assumptions and rely on an *hermeneutical horizon* – an intuitive perception what concepts and assumptions are basic and true and how we understand them. Thus, any ontology will achieve the ideal goals mentioned above only to a certain degree. Note, however, that this implies that any ontology can be re-engineered, corrected according to changes in the hermeneutical horizon.

Thus, known ways of constructing ontologies can be treated not as absolute recipes, but hints how to proceed. There is a distinction of a top-down approach - actually, starting with an intuitive perception of the basic concepts in hermeneutical horizon and specifying them in detail subsequently – and a bottom-up approach - starting, say, with the concepts actually used in a given field of knowledge and trying to interpret them and their structural relations. The top-down approach starts with issues related to meta-model functionality (idea goals, 5); the bottom-up approach starts with issues related to systematization (ideal goals, 3) and standardization (ideal goals, 4). Obviously, we need a combination of both approaches in order to construct a useful ontology.

In the paper, we tried to construct the ontology of 21st Century COE Program *Technology Creation Based on Knowledge Science* at JAIST as a case study, with the following goals:

(1) To clarify the use of the concept of Knowledge Science in this Program and make explicit (at least, as much as possible) assumptions about this concept that are often tacitly made (ideal goals 2, 5);

(2) To represent a vocabulary of terms used in this COE Program, together with a systematization of terms used (ideal goals 1, 3);

(3)To help in the development of a software system designed to support hermeneutic search of literature, and possibly in other projects related to the COE Program.

The ideal goal 4) – standardization – is addressed only to limited degree, because of the heterogeneity of the interdisciplinary projects in the COE Program. Thus, we design ontology for COE program at JAIST not only for helping in the development of some projects of this program, but also make to clarify basic concepts for COE program itself.

To create ontology, we proceeded along several lines. First, we checked the terms and concepts used by the program leader in a paper presenting an introduction to the COE program, thus providing an top-down outline of COE ontology. Then, we collected 43 papers composed by COE project members, which have appeared either at an international conference or journal. We extracted the keywords from the papers and counted the frequency of keywords in the full paper by using a computer program. We chose the keywords with high frequency to supplement the outline of COE ontology. We chose also pairs of keywords occurring with non-zero frequency to make a simple QT clustering of them [4] and compared the ontology emergent bottom-up from such clustering with the top-down outline of COE ontology. Finally, we took into account the reflection on knowledge sciences presented in a recent paper [5] and used this reflection for corrections of the supplemented outline; this way, we finally created the ontology for COE program. To understand our result better, we reflected also about other possible views on such ontology and analyzed at least one possible application of the ontology.

2. Bottom-Up Classification and Specification: Keyword Analysis

To build an outline of the ontology of COE program, we started with the paper presenting an introduction to this program authored by the program leader [6]. After analyzing the purpose and sub-projects of the program, we selected the key terms and concepts mentioned in the paper and organized an ontology outline with three levels of branches. The first level included five main topics:

- Knowledge science,
- Systems science and methodology,
- Education in knowledge science,
- Knowledge creation,
- Management of technology.

In addition, we also referred to the program reports presented by the program leader in later periods to check and revise the outline.

Furthermore, we collected the papers authored by COE project members - as many as were available. Since we had to limit this search to electronic files, we finally considered only 43 papers, which were either included in Proceedings of International Symposium on Knowledge and Systems Sciences (JAIST, 2004), or Proceedings of the First World Congress of the International Federation for Systems Research (Kobe, 2005), or in the International Journal of Knowledge and Systems Sciences (Issues 1 to 6). We extracted the keywords from all papers and counted the frequency of their occurrences in the full body of papers by using a computer program designed by a member of our group.

2.1 Keywords extraction

We consider that keywords specified by authors are scarce. An additional keyphrase extraction can enlarge the set of keyphrases for each paper. This may increase the correlation and improve the clustering of the keyphrases. Keyphrase extraction techniques for the English language are known. They involve the use of regular expressions to identify parts of speech, and then indentify the keyphrases which are composed of the certain parts of speech, for example, an adjective and a noun, or two nouns one after another: "knowledge management", "information retrieval".

We also recognized that some high fequency words, such as the closed-class words — the, a, an, in, to, were — too common to be significant. We set up a high cutoff called "stop words" which filtered out high frequency common words, and a low cutoff which eliminated insignifcant low frequency words. Words between these two cutoffs we considered as possessing "resolution power" (the ability of words to discriminate text contents). The two cutoffs were determined experimentally.

Based on the knowledge bases of "*stop words*", there are three steps to the extractor algorithm:

(1) *Find Keyphrases*: Extract keyphrases from the text file and make a list of all phrases. A phrase is defined as a sequence of one, two, or three words that appear consecutively in the text, with no intervening stop words or punctuation. In our case, phrases of four or more words are relatively rare. Therefore Extractor only considers phrases of one, two, or three words.

(2) Score Keyphrases: For each keyphrase, count how often the keyphrase appears in the text. Assign a score to each phrase. The score is the number of times the keyphrase appears in the file.

(3) Final Output: We now have an ordered list of mixed-case phrases (upper and lower case, if appropriate). The list is ordered by the scores calculated in step 2.

2.2 Keywords clustering

Another attempt was a clustering of keywords based on their joint occurrence. We selected a simple QT (quality threshold) clustering algorithm [4]. The goal of OT clustering is go form large clusters of genes with similar expression pattern, and to ensure a quality guarantee for each cluster. Quality is defined by the cluster diameter and the minimum number of genes contained in each cluster. In our case, if the frequency of occurrence of a pair of keywords equals or exceeds an assumed threshold t, the pair might be counted to belong to a candidate cluster; the largest of such candidate clusters is counted as an actual cluster, it is subtracted from the entire set of keywords, and the procedure is repeated on the remaining keywords. It turned out that the joint occurrence of keywords is not common, most frequencies of such co-occurrence are zero, thus the clustering was done at the threshold level $t = 1$. Because of the space, we only addressed the output of three clusters:

COE Program papers:
Keywords (genes) : 128
Threshold: t = 1

Cluster1:

Papers:
{
04_1_tianjin.txt, 04_3_Saito-Medeni-Machado_v2.txt, 20068.pdf.txt, 20017.pdf.txt, 20038.pdf.txt, 20063.pdf.txt, 20068.pdf.txt, 20219.pdf.txt, 20074.pdf.txt
}
Keywords:
{
Scientific knowledge creation, Knowledge management, Knowledge management system, I-system, Laboratory knowledge management, Knowledge management education, Curriculum development, Degree programs, Knowledge science, Systems concepts, Creative environments, Organizational knowledge creation, Workflow for process analyses, Knowledge-creating process, Soft system methodology, Concept creation, Pattern of innovation
}

Cluster2:
Papers:
{
09_1_Minh.txt, 09_2_Nagai-kss04.txt, 12_2_phan.txt, 12_3_Tran.txt, 15_1_Zhang.txt, 15_2_huang-wei.txt, 20055.pdf.txt, 20057.pdf.txt, 20073.pdf.txt, 06_1_Hao.txt, 20177.pdf.txt
}
Keywords
{
Cross language text summarization, Text summarization, Natural language processing, Text mining, Association rule mining, Coreference resolution, Anaphora resolution, Clustering algorithm, Information extraction, Natural language processing, Data mining, Knowledge discovery, Clustering, Genetic algorithm, K-means algorithm, Text clustering, Ant-based Clustering, Semantic similarity measure, Ontology, Phrase indexing, Text summarization, Sentence extraction, Ensemble learning, SVM ensemble, Direct space method, Rough sets
}

Cluster3:
Papers:
{

05_1_ma.txt, 05_3_ JieYAN.txt, 20060.pdf.txt
}
Keywords:
{
Transportation fuel cell forecast, Technology roadmapping, Systems thinking, Technology creation, Systemic thinking, Roadmapping process, Technology forecasting, Roadmapping, Interactive planning
}

With respect to the outline of COE ontology with three levels branch that we summarized from the project introduction and reports, the key phrases included in cluster one belong to the topics “Knowledge Science”, “Knowledge Creation” and “Education in Knowledge Science”; the key phrases included in cluster two belong to the topic of “Knowledge Representation and Acquisition”, we have not listed a independent branch for it yet; the key phrases included in cluster three belong to the topic of “Management of Technology” and “Systems Science and Methodology”. Thus, these clusters give us the hints to category the keywords, rethink the ontology outline of COE Program and finally construct a program ontology from the bottom-up method.

3. Top-Down Approach: a Reflection on the Concept of Knowledge Science

Knowledge science (KS) is often confused with or tacitly assumed to be subordinated to *knowledge management* (KM), thus we first reflect on the origins and meaning of the second term. *Knowledge management* has much popularity in management science, but its technological origins are often forgotten. It was first introduced by computer technology firms in early 1980-ies – first in IBM, then Digital Equipment Corporation who probably was the first to use the term *knowledge management* – as a computer software technology in order to record the current work on software projects. This started the tradition of treating knowledge management as a system of computer technologies. Later this term was adopted by management science, and made a big career. This has led to two opposite views how to interpret this term [7][8]:

- As *management of information relevant for knowledge-intensive activities*, with stress on information technology: databases, data warehouses, data mining, groupware, information systems, etc.
- As *management of knowledge related processes*, with stress on organizational theory, learning, types of knowledge and knowledge creation processes.

The first view is naturally represented by information technologists and hard scientists; the second by social scientists, philosophers, psychologists and is clearly dominating in management science. Representatives of the second view often accuse the first view of perceiving *knowledge to be an object* while it should be seen as *knowledge related to processes;* they stress that knowledge management should be *management of people*. For example, in an excellent book on the dangers of postponing action *The Knowing-Doing Gap* [9], say that “[an] article asserted that ‘knowledge management starts with technology’. We believe that this is precisely wrong. …Dumping technology on a problem is rarely an effective solution.”

However, while it is correct that knowledge management cannot be reduced to management of information, such *a correct assessment is a pitfall* (an unfortunate impact of binary logic on our thinking): being sure that they are right, the representatives of the second view overlook both the complexity and the essence of the controversy. The complexity is that, historically, knowledge management has started with technology and cannot continue without technology; thus, both interpretations should be combined in adequate proportions. The essence of the controversy is that *management of people* should be also understood as *management of knowledge workers;* and knowledge workers are today often mostly information technologists, who should be well understood by managers. Thus, we believe that the two views listed above should be combined. Moreover, they incompletely describe what knowledge management is; there is a third, essential view, seeing knowledge management as the *management of human resources in knowledge civilization era,* concentrating on knowledge workers, their education and qualities, assuming a proper understanding of their diverse character, including a proper understanding of technologists and technology.

This is particularly visible concerning the concepts of *technology management* versus *knowledge management.* Management science specialists in knowledge management often tend to assume that *technology management* is just a branch of *knowledge management;* technologists specializing in *technology management* stress two aspects. However, an essential meaning of the word *technology* is *the art of designing and constructing tools or technological artefacts* (thus, *technology* does not mean *technological artefacts*, although such a meaning is often implied by a disdainful use of the word *technology* by social sciences, e.g., in the quoted above phrase *dumping technology*). In this sense, the term is used in the phrase *technology management.* Secondly, *technology management* might be counted as a kind of special *knowledge management*, but it is an older discipline, using well developed concepts and processes, such as *technology assessment, technology foresight* [10] and *technology roadmapping* [11][12]. Only recently, some of these processes have been also adapted to knowledge management [13].

All the above discussion implies that we are observing now an emergence process of a new understanding of *knowledge sciences* – an interdisciplinary field that goes beyond the classical epistemology, includes also some aspects of *knowledge engineering* from information technology, some aspects of *knowledge management* from management and social science, some aspects of *interdisciplinary synthesis* and other techniques (such as decision analysis and support, multiple criteria analysis, etc.) from systems science. This emergence process is motivated primarily by the needs of an adequate education of *knowledge workers* and *knowledge managers and coordinators;* however, also the research on knowledge and technology management and creation needs such interdisciplinary support.

The classical understanding of the words *knowledge science* might imply that it is epistemology enhanced by elements of knowledge engineering, knowledge management and systems science. However, the strong disciplinary and historical focus of epistemology suggests an opposite interpretation: knowledge science must be interdisciplinary, thus it should not start with epistemology, although it must be enhanced by elements of epistemology. The field closest to knowledge science seems to be systems science – at least, if it adheres to its interdisciplinary origins and does not suffer too much from the unfortunate (but unavoidable today) disciplinary division into *soft* and *hard systems science.* The noticeable tension between *soft* and *hard* systems science is just an older version of the tension between understanding *knowledge management* either from the perspective of knowledge engineering, or from the perspective of social and management science, mentioned above.

To summarize, we should thus require that *knowledge sciences* gives home to several disciplines (quoted here in an alphabetic order):

- *Epistemology and philosophy of science,*
- *Knowledge engineering,*
- *Management science and knowledge management,*
- *Sociological and soft systems science,*
- *Technological and hard systems science,*

To our knowledge, only one university in the world, the Japan Advanced Institute of Science and technology, founded – already in 1998 – the School of Knowledge Science, while the field is understood similarly as described above. The university supports only graduate education, for master and doctoral degrees; in knowledge science, three types of graduates are typical:

- Specialists in management, with understanding of knowledge engineering and systems science;
- Specialists in systemic knowledge coordination, with understanding of knowledge engineering and management;
- Specialists in knowledge engineering, with understanding of management and systems science.

On equal footing, with a requirement of mutual information and understanding, this basic classification should be also reflected in the proposed ontology of the COE Program.

4. Final Proposal of the Ontology and its Application

Based both on the bottom-up classification and on the above reflection as a basis of top-down approach, the ontology of the COE Program can be proposed. It is organized as an inverted tree, with fourth-level branches corresponding to keywords found in the papers of COE Program members. The general category of the domain of Knowledge Science includes the following eight sub-domains as the first lever of ontology of the COE Program:

- *Knowledge Creation and Transformation*
- *Knowledge Representation, Systematiza-*

tion, Acquisition

- *Knowledge Management*
- *Systems Science*
- *Education and Knowledge Science*
- *Management of Technology*
- *Technology Creation*
- *Diverse Related Themes*

Each sub-domain is consisted of several topics (Second lever); the different topics include particular sub-topics (Third Lever). All keywords was summarized as and categorized into the sub-topics (Fourth lever). In addition, the clustering of the keywords gave us the hints to find the relations between the subtopics and the further relations between topics as well as sub-domains. The structure of the ontology is not only a simple tree, also a network. Because of the limitation of pages, I can not list the proposed ontology here. Our classification is naturally not absolute nor the ultimately final; it might be further enhanced and corrected as new data will become available.

On the basis of requirements of researchers [14][15] and the phenomenon of *Hermeneutics*[16], a software tool for information and knowledge retrieval was designed [17][18], in order to help researchers in gathering and interpreting relevant knowledge or research materials; this software tool is called *Adaptive Hermeneutic Agent (AHA)*. The AHA is equipped with a simple and intuitive search interface and uses familiar search syntax, such as used by popular search engines (like Google, Yahoo). The search support can be extended to the definition of queries that will be automatically executed by the system with a fixed period of time. The definition of a query by the user is helped by ontological information; actually, the ontology described above is used in AHA as a basis of defining queries that can be selected from this ontology, supplemented or modified, for example, by adding new keywords that are relevant to the searched topic. After the query is executed, the AHA can also filter the obtained results by using a reinforcement learning approach that relies on a profile of the user's interests. The AHA could also use a visual interface for the clustering and graphical presentation of search results.

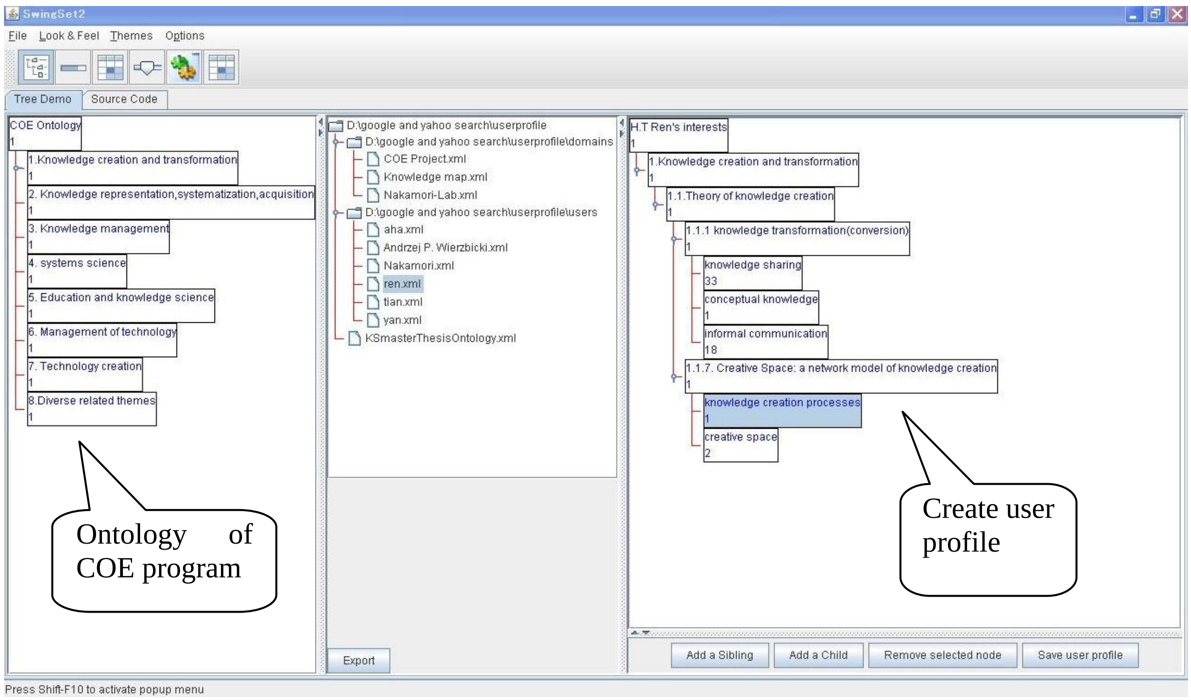

Figure 1. The main interface of creating user profile based on ontology of COE Program

Therefore, the COE ontology as described earlier is an important element, first step in developing the software tool of AHA. The second step is the creation of user profile. The user, for example, a COE member, could extract the knowledge from COE ontology to formulate the outline of user profile, for example, select the domains (keywords) he are most interested in and give the weights for different keywords. Then, the user could gather relevant knowledge and information based on his profile by using search engines connected to AHA. The AHA will do adaptive selection automatically as following steps: text extraction (from MS-word file to text or from PDF file to text); keyword extraction and frequents calculation (extracting keywords from the search results by statistics method); measurement of the similarity of each file and user profile; giving a ranking list including top N results. The figure 1 shows a interface of creating user profile based on ontology of COE Program.

Other possible applications of the work on ontology formation described here include, for instance, the development of an ontology of Knowledge Science in JAIST, an ongoing project that will include the lessons from the work described here; or a construction of a Knowledge Map or a research network for professionals interested in related domain, etc.

5. Conclusions

We presented a process of constructing ontology of the 21st Century COE Program *Technology Creation Based on Knowledge Science* together with one of possible applications – helping in the development of an adaptive hermeneutic agent (AHA). The construction of ontology is a complex, multidimensional process; we must combine bottom-up approaches (from recorded documents) with top-down processes (from intuitive hermeneutical horizon), also look from diverse perspectives to improve the final product. Nevertheless, the effort spent on ontology construction is profitable in terms of diverse possible applications and of a creative illumination and enlightenment.

Acknowledgment

The research is supported by 21st COE Program *Technology Creation Based on Knowledge and Science* of JAIST, a fund by the Ministry of Education, Culture, Sports, Science and Technology (MEXT, Japan).

References

[1] Heidegger M.. *Sein und Zeit.* Niemayer, Halle, 1927

[2] Dieng R., Corby O., *Knowledge Engineering and Knowledge Management: Methods, Modesl and Tools.* Springer Verlag, Berlin-Heidelberg, 2000.

[3] Król Z., The Emergence of Complex Concepts in Science. In Wierzbicki A.P. and Nakamori Y., eds.: *Creative Environments: Issues of Creativity Support for the Knowledge Civilization Age.* Springer Verlag, Berlin-Heidelberg 2007, in print.

[4] Heyer L.J., Kruglyak S, Yooseph S., Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.*, 9, 1106–1115, 1999.

[5] Wierzbicki A.P., Nakamori Y., Knowledge sciences – some new developments. *Zeitschrift für Betriebswirtschaft*, 2007, in print.

[6] Nakamori Y., Introduction to a COE Program at JAIST. Proceedings of International Forum ‘Technology Creation Based on Knowledge Science: Theory and Practice’, Nov. 10-12, JAIST, Japan, pp. 1-4, 2004.

[7] Wiig K.M., Knowledge management: an introduction and perspective. *Journal of Knowledge Management* 1(1): 145-156, 1997.

[8] Davenport T. and Prusak L., *Working Knowledge: How Organizations Manage What They Know.* Harvard Business School Press, Boston Ma, 1998.

[9] Pfeffer J., Sutton R.I.: *The Knowing – Doing Gap: How Smart Companies Turn Knowledge Into Action.* Harvard Business School Press, Boston, Ma, 2000.

[10]Salo A., Cuhls K. Technology Foresight – Past and Future. *Journal of Forecasting* 22(2-3):79-82, 2003.

[11]Willyard C.H., McClees C.W., Motorola’s technology roadmap process. *Research Management* 30(5):13-19, 1987.

[12]Phaal R., Farrukh C., Probert D., Technology roadmappingg - a planning framework for evolution and revolution. *Technological Forecasting and Social Change* 71:5-26, 2004.

[13]Ma T., Liu S., Nakamori Y., Roadmapping

for Supporting Scientific Research. *Proceedings of MCDM 2004* (17th International Conference on Multiple Criteria Decision Making, Whistler, Canada), Simon Fraser University.

[14]Tian J., Nakamori Y., Knowledge Management in Scientific Laboratories: A Survey-based Study of a Research Institute. *Proceedings of the Second International Symposium on Knowledge Management for Strategic Creation of Technology*, 19-26, 2005.

[15]Tian J., Wierzbicki A.P., Ren H. and Nakamori Y., "A Study of Knowledge Creation Support in a Japanese Research Institute", *International Journal of Knowledge and System Science*, Vol. 3, No. 1, pp.7-17, March 2006.

[16]Wierzbicki, A.P. and Nakamori Y., *Creative Space: Models of Creative Processes for the Knowledge Civilization Age.* Springer Verlag, Berlin-Heidelberg 2006.

[17]Ren H., Wierzbicki A.W.. Implementing Creative Environments for Scientifc Research. Journal of Systems Science and Systems Engineering, to be published, 2007.

[18]Ren H., Tian J., Nakamori Y., Wierzbicki A.P. Electronic Support for Knowledge Creation in a Research Institute. J*ournal of Systems Science and Systems Engineering*, to be published, 2007.