ECAI 2012
20TH EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE
MONTPELLIER, FRANCE
AUGUST 27-31, 2012

ifip

Proceedings

1st International Workshop on

Artificial Intelligence for Knowledge Management (AI4KM 2012)

August, 28th, 2012, Montpellier, France

Eunika Mercier-Laurent
Nada Matta
Mieczyslaw L. Owoc
Ines Saad
Editors

Knowledge management is a large multidisciplinary field having its roots in Management and Artificial Intelligence. AI brought the way of thinking, knowledge modelling, knowledge processing and problem solving techniques.

Knowledge is one of intangible capitals that influence the performance of organizations and their capacity to innovate. Since the beginning of the KM movement in the early nineties, companies and non profit organizations have been experiment various approaches, often without using AI.

The objective of this multidisciplinary workshop is to gather both researchers and practitioners to discuss methodological, technical and organizational aspects of AI used for knowledge management and to share the feedback on KM applications using AI.

AI4KM 2012 Program

8:30 Opening session
Structured or natural knowledge representation for KM: 30 years of compromises between humans and machines, Jean Rohmer, Pole Universitaire Leonard de Vinci.

Lecture (in proceedings):
Overview of AI path in USSR and Ukraine. Up-to-date Just-In-Time Knowledge Concept
K.M. Golubev, General Knowledge Machine Research Group

9:10 Session 1
Session chairs E. Mercier-Laurent Univ Lyon3, G. Kayakutlu ITU

9:10 *From Community Memories to Corporate Memory* - Daniel Galarreta and Pascale Riviere, CNES.
9:30. *Contextual knowledge handled by an expert* - Janina Jakubczyc and Mieczyslaw Owoc,
9:45.*Artificial Intelligence for Knowledge Management with BPMN and Rules* - Antoni Ligęza, Krzysztof Kluza, Grzegorz J. Nalepa and Tomasz Potempa
10:05 *Discussion*

10:30 *Coffee break*

10:45 Session 2
Session chairs E. Mercier-Laurent Univ Lyon3, Antoni Ligęza, AGH

10:45. *Web User Navigation Patterns Discovery as Knowledge Validation challenge* Pawel Weichbrot and Mieczyslaw Owoc
11:00 *Kleenks:collaborative links in the Web of Data* Razvan Dinu, Andrei-Adnan Ismail, Tiberiu Stratulat and Jacques Ferber
11:20 *From Knowledge transmission to Sign sharing: Semiotic Web as a new paradigm for Teaching and Learning in the Future Internet* Noel Conruyt, Véronique Sebastien, Didier Sébastien, Olivier Sebastien and David Grosser.
11:40 Discussion

12: 00 lunch

13:40 Session 3
Session chairs E. Mercier-Laurent Univ Lyon3, Mieczyslaw L.Owoc, AE

13 :40 *Ontology Learning From Unstructured Data for Knowledge Management: A Literature Review* Jens Obermann and Andreas Scheuermann
14 :00 *Description Logic Reasoning in an Ontology-Based System for Citizen Safety in Urban Environment*, Weronika T. Adrian, Antoni Ligęza and Grzegorz J. Nalepa.
14:20 *Advanced System forAcquisition and Knowledge Management in Cultural Heritage* Stefan Du Chateau, Danielle Boulanger and Eunika Mercier-Laurent.
14 :40 *GAMELAN: A Knowledge Management Approach for Digital Audio Production Workflows* Karim Barkati, Alain Bonardi, Antoine Vincent and Francis Rousseaux
14:55 *Discussion*

15:15 *coffee break*

15:30 Session 4
Session chairs E. Mercier-Laurent Univ Lyon3, Noel Conruyt Univ Reunion

15:30 *Knowledge Management applied to Electronic Public Procurement* Helena Lindskog, Eunika Mercier-Laurent and Danielle Boulanger.
15:40 Gulgun Kayakutlu and Ayca Altay. *Collective intelligence for Evaluating Synergy in Collaborative Innovation*
16 :00 *Discussion*
16 :15 *Closing discussion – AI for KM Challenges*

AI4KM 2012 Invited talk

Jean Rohmer - Pole Universitaire Leonard de Vinci

Structured or natural knowledge representation for KM: 30 years of compromises between humans and machines

Abstract:

If we want to provide knowledge management with some artificial intelligence, we must achieve contradictory objectives: let humans interact with knowledge in a natural, acceptable format -not through formal notations. In another hand, machines can conduct automatic reasoning only if they operate on a precise, computable structuration of information. We will explore the various proposals in both directions over the past years, and contemplate some tracks for the future, taking into account modern hardware, software and practices, like social networks. We eventually will propose our own compromise, namely "litteratus calculus"

Overview of AI path in USSR and Ukraine. Up-to-date Just-In-Time Knowledge Concept

K.M. Golubev
General Knowledge Machine Research Group

Abstract. This paper contains a short description of AI history in USSR and Ukraine. It describes also a state-of-the-art approach to intellectual activity support called Adaptive Learning based on the Just-In-Time Knowledge concept. It's kind of the Artificial Intelligence and Knowledge Management fusion. To obtain more detailed information, please visit site *http://gkm-ekp.sf.net*.

Keywords. adaptive learning, just-in-time knowledge, general knowledge machine, electronic knowledge publishing

Introduction

The history of computing and AI at particular in USSR is not widely known. The author tries to tell about remarkable people and ideas born in Soviet times in a hope that it could be interesting and inspiring. Information is derived from sources publicly accessible.

The author tries also to tell about AI path in the post-Soviet times in Russia and Ukraine. The original approach to knowledge presentation and learning developed by General Knowledge Machine Group based in Kiev, Ukraine, is described.

1. Remembering a remarkable Soviet computing pioneer

http://googleblog.blogspot.com/2011/12/remembering-remarkable-soviet-computing.html

From Google Official Blog:

"December 25, 2011

In many parts of the world, today is Christmas—but in Russia and Eastern Europe, which use the Orthodox calendar, December 25 is just an ordinary day. Little known to most, however, it's also a day that marks the anniversary of a key development in European computer history.

Sixty years ago today, in the Ukrainian capital of Kyiv, the Soviet Academy of Sciences finally granted formal recognition to Sergey Lebedev's pioneering MESM project. MESM, a Russian abbreviation for "Small Electronic Calculating Machine," is regarded as the earliest, fully operational electronic computer in the Soviet Union—and indeed continental Europe.

Recently we were privileged to get a first-hand account of Lebedev's achievements from Boris Malinovsky, who worked on MESM and is now a leading expert on Soviet-era computing.

Described by some as the "Soviet Alan Turing," Sergey Lebedev had been thinking about computing as far back as the 1930's, until interrupted by war. In 1946 he was made director of Kyiv's Institute of Electrical Engineering. Soon after, stories of "electronic brains" in the West began to circulate and his interest in computing revived.

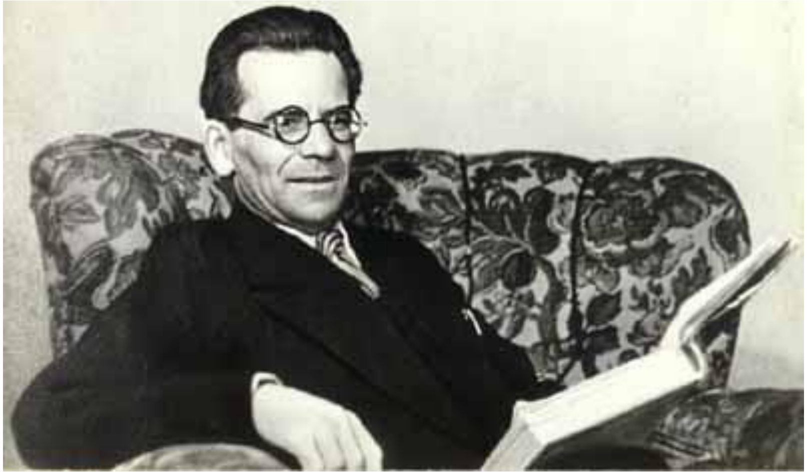


Figure 1. Sergey Lebedev*

Initially, Lebedev's superiors were skeptical, and some in his team felt working on a "calculator"—how they thought of a computer—was a step backward compared to electrical and space systems research. Lebedev pressed on regardless, eventually finding funding from the Rocketry department and space to work in a derelict former monastery in Feofania, on the outskirts of Kyiv.

Work on MESM got going properly at the end of 1948 and, considering the challenges, the rate of progress was remarkable. Ukraine was still struggling to recover from the devastation of its occupation during WWII, and many of Kyiv's buildings lay in ruins. The monastery in Feofania was among the buildings destroyed during the war, so the MESM team had to build their working quarters from scratch—the laboratory, metalworking shop, even the power station that would provide electricity. Although small—just 20 people—the team was extraordinarily committed. They worked in shifts 24 hours a day, and many lived in rooms above the laboratory.

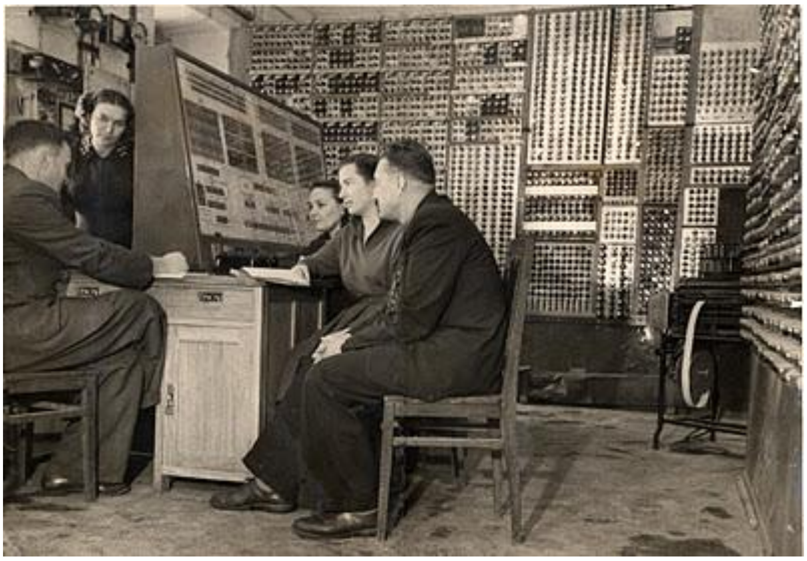


Figure 2. MESM and team members in 1951. From left to right: Lev Dashevsky, Zoya Zorina-Rapota, Lidiya Abalyshnikova, Tamara Petsukh, Evgeniy Dedeshko

MESM ran its first program on November 6, 1950, and went into full-time operation in 1951. In 1952, MESM was used for top-secret calculations relating to rocketry and nuclear bombs, and continued to aid the Institute's research right up to 1957. By then, Lebedev had moved to Moscow to lead the construction of the next generation of Soviet supercomputers, cementing his place as a giant of European computing. As for MESM, it met a more prosaic fate—broken into parts and studied by engineering students in the labs at Kyiv's Polytechnic Institute.

**All photos thanks to ukrainiancomputing.org.*

Posted by Marina Tarasova, Communications Associate, Ukraine"

2. Victor Glushkov – Institute of Cybernetics Founder

Figure 3. Victor Glushkov

Victor Glushkov (August 24, 1923 – January 30, 1982) was the founding father of information technology in the Soviet Union (and specifically in Ukraine), and one of the founders of Cybernetics. He was born in Rostov-on-Don, Russian SFSR, in the family of a mining engineer. He graduated from Rostov State University in 1948, and in 1952 proposed solutions to the Hilbert's fifth problem and defended his thesis in Moscow State University.

In 1956 he began working in computer science and worked in Kiev as a Director of the Computational Center of the Academy of Science of Ukraine. In 1958 he became a member of the Communist Party.

He greatly influenced many other fields of theoretical computer science (including the theory of programming and artificial intelligence) as well as its applications in USSR.

Glushkov founded a Kiev-based Chair of Theoretical Cybernetics and Methods of Optimal Control at the Moscow Institute of Physics and Technology in 1967 and a Chair of Theoretical Cybernetics at Kiev State University in 1969. The Institute of Cybernetics of National Academy of Science of Ukraine, which he created in 1962, is named after him.

3. Nikolay Amosov - Founder of Bio-cybernetic Information Technologies

“The sphere of interests of outstanding surgeon Nikolai Mihailovich Amosov included not only medical problems, but also general human cognition problems. General system approach to understanding human nature have been reflected in the scientific directions initiated by N.M.Amosov in various areas of cybernetics: modeling of physiological functions of human organism (physiological biocybernetics), modeling of cognitive and psychological human functions (psychological biocybernetics), modeling a man as a social creature (sociological biocybernetics). All these research directions have been represented in the Department of Biocybernetics founded in the Institute of Cybernetics by V.M.Glushkov and N.M.Amosov in 1961. Nikolai Mikhailovitch Amosov was the scientific leader of the Department since 1988.

In 1964 Nikolai Mikhailovitch Amosov formulated a hypothesis on the information processing mechanisms of the human brain. Within this hypothesis he expressed his system-level observations on the brain's structure and the mechanisms that are made operational by a human's mental functions. Of principal importance was the fact that it was not the separate structures, mechanisms or functions (such as memory, perception, learning and so on) that became the simulation object, but the brain of the human as a social being - the brain of homo sapiens. Such was the main idea of the monograph "Modeling of Thinking and of the Mind ", published in 1965, which for a couple of decades became the bible for several generations of Department's researchers (and not only for them).

The ideas, which N.M.Amosov put forward in his book "Modeling of Thinking and of the Mind " were further developed in his subsequent works ("Modeling of Complex Systems ", "Artificial Intelligence ", "Algorithms of the Mind ", "Human Nature ").

On a theoretical level, two main features characterize the research of Amosov School.

The first feature is that not an individual neuron, but a set of neurons organized in a particular way - neuron assembly - is considered to be the core functional element of a neural network, its "principal character ". Given this, the neural network appears now as a structure consisting of a multitude of interacting assemblies, each of which corresponds (and this is a very important point) to some individual image or concept out of a set of images and concepts that participates in forming integrative mental functions realized by the brain. That is, this set participates in the thought process. Thus the neural network turns out to be a network with semantics (a special kind of a semantic network). The origins of the present approach can be traced to the early works of a well-known physiologist D.Hebb, whose main study was published as early as in 1949.An important characteristic of this kind of network is that all of its elements at any point in time are active to some degree. The magnitude of this activity varies in time, reflecting the interaction of concepts represented by the network's nodes.

The second feature of Amosov school research concerns the introduction of the notion of a specific system for reinforcement and inhibition (SRI) to scientific use. This system is an integral neural network attribute, and in network functioning it plays a role comparable to that of functions of attention in the thought processes. The idea of SRI is entirely original. Using this system allows to introduce a direction component into neural network information processing, and, what is very important, to use value characteristics of information in organizing this processing.”

Dr.Alexander Kasatkin, Dr.Lora Kasatkina

International Research and Training Center of Information Technologies and Systems of National Academia of Sciences of the Ukraine

4. Dmitry Pospelov- Founder of Russian Artificial Intelligence Association

Dmitry Pospelov was born on 19.XII.1932 , Moscow.

Technical Sciences Doctor, Professor, Member of Russian Academy of Natural Sciences (10.X.1990).

Graduated from Lomonosov Moscow State University, as a Computational Mathematics specialist. Head of Artificial Intelligence Problems Department at the Computer Center of Russian Academy of Sciences named after A.A. Dorodnitsin.

Head of International UNESCO Artificial Intelligence Laboratory. Head of "Intellectual Systems" division of Russian Academy of Sciences.

Founder of Russian(Soviet) Artificial Intelligence Association (www.raai.org).

From: From Universal Scales to Systems of Communicating Contextual Systems by Irina Ezhkova

http://posp.raai.org/data/posp2005/Ezhkova/ezhkova.html

"It was more than four decades ago when Dmitry Pospelov began his inspiring study of the Semiotic Systems, Situated Logics, Universal Scales and Spaces. His interest in psychology and neurology, in mathematical logics and fuzzy sets, in linguistics and behavior sciences had been stimulated the blossoming tree of a broad Russian school of AI. His typical way of approaching constructive model was formalized as a cortege, or train (or even a simple list) of elements, each of which then may be represented well in a traditional way. This reflected his original flexibility, profound vision and interdisciplinary views.

His intuition was deeply based on a belief that Osgood scales and related spaces may lead to a better understanding of semantically grounded systems. This finally has lead to a discovery and development of the Universal Scales. Latter research in this direction allowed development of the unified integrating framework for modeling a diversity of cognitive and complex real phenomenon. The basic principles of Cognitive Relativity, Rationality and Clarity were crystallized to underline this direction of the Russian school of thought. It became clear that both views can be integrated on the basis of these principles. The mathematical theory of Systems of Communicating Contextual Systems is based on recursive mechanisms of theorem proving and constraints recognition and satisfaction, the first elements of which were also developed in 1974-1978 under the supervision of Dmitry Pospelov, and in a productive collaboration with other Russian mathematical schools such as of Prof. Maslov and of Prof. Kotov.

The Contextual theory of Cognitive States and the Systems of Communicating Contextual Systems (C2S) suggest a unified framework for modeling life-cycles of patterns, representations, and of possible ways of their construction, generation, interaction and transformation. This framework allows modeling complex center-activated or distributed self-organizing phenomenon, which may have centered or distributed cognition. It allows invention of a new kind of AI systems, Evolutionary Evolving Intelligent Systems (EI), which are based on what we call by λ-Intelligence, and which are principally open and flexible, continuously learning, self organized, cognitively tailored and collectively adaptive systems ".

5. Mikhail Livanov - Spatial Organization of Cerebral Processes Research

http://www.amazon.com/Spatial-Organization-Cerebral-Processes-Livanov/dp/0706515145/ref=la_B001HPVRUY_1_1?ie=UTF8&qid=1343818847&sr=1-1

Born on 7(20).10.1907, Kazan, Soviet physiologist, member of USSR Academy of Sciences (1970). Graduated from Kazan State University (1931). Head of laboratory at Institute of Higher Nervous Activity and Neurophysiology of Academy of Sciences

From: [The phenomenon of spatial synchronization of the brain potentials in a broad frequency band 1-250 Hz]

http://lib.bioinfo.pl/paper:18064890

“The article dedicated to the centenary of academician Mikhail Nikolaevich Livanov briefly outlines the history of development of his original concept of the functional significance of the brain potential's spatial synchronization phenomenon as a possible way of studying systemic organization of the brain electrical activity. The new parameter of "space" introduced into neurophysiology by M. N. Livanov made it possible to research the earlier unknown aspect of the brain activity. Livanov's ideas have been developed in many studies of the late decades of the XX century. In the review, much attention is given to specific functional significance of this phenomenon in a broad frequency band 1-250 Hz, especially, during instrumental learning. Energy (power spectra) and coherent-phase characteristics of cortical potentials in traditional (1-30 Hz), gamma-(30-80 Hz) and high-frequency (80-250 Hz) bands are compared. The problem of linear and nonlinear processes in the organization of the brain potentials is mentioned.”

6. General Knowledge Machine Research Group

General Knowledge Machine Research Group was founded in 1986 in Kiev, Ukraine, as informal institution by mathematicians and IT experts. It counts 11 members including sponsors, developers and thinkers. The author of paper has generated initial ideas, coordinates activities and plays all roles needed for the show.

6.1. LEARNING

Exams:

Prof.: You are looking very worried. Any problems with exams questions?
Stud.: Oh, no! Questions are OK. It is the answers that I worry about.

6.2. Traditional learning

Traditional learning is based on a linear process, when students must learn all proposed knowledge, topic by topic. After that students must pass exams to get acknowledgement from professors that knowledge is in their minds. Initial time of learning is very big, usually up to 17 years (school + university). There are many exams, sometimes very difficult, having significant influence on the life of students.

But all this very hard work does not guarantee that students have all or even greater part of knowledge needed to solve problems which arise in their post-school activity, in the real world life.

6.3. Adaptive Learning

Adaptive Learning is based on a concept called Just- In-Time Knowledge (JIT-Knowledge). Total amount of external sources of knowledge, even in the specific areas, becomes greater all the time. It is not possible, taking into account limitations of human brain, to learn it with Traditional Learning, topic by topic. It means that in reality significant part of knowledge is not used by anyone, and many problems are not solved because no one learns needed knowledge. The Electronic Knowledge Publishing based on General Knowledge Machine power, called GKM-EKP technology, allows to find and to learn knowledge relevant to existing problems.

Table 1. Comparison of AI Expert Systems and Electronic Knowledge Systems

AI Expert Systems	Electronic Knowledge Systems
Intended to replace human experts	Intended to assist human intellect
Based primarily on mathematics	Based on neurophysiology, psychology, knowledge management theory and mathematics
It is practically impossible to transform directly external knowledge sources to expert systems	It is further advancement of a traditional publishing – external knowledge sources (books, articles etc) may be transformed into e-knowledge systems .
Based on the decision rules concept	Based on the general knowledge concept using approach developed by Academician of USSR M.N.Livanov
It is relatively hard work to incorporate an expert system into other information systems due to sequential nature of data input and output	E-knowledge system may be easily incorporated into any kind of information system due to support of wide range of data input and output sources
It is practically impossible to use expert systems for learning, because they are not based on human knowledge	It may be used for Adaptive Learning applications, based on the Just-In-Time Knowledge concept

6.4. Steps of intellectual activity

Following Mr. Sherlock Holmes, we can describe steps of expert’s activity:

- Observation
- Producing propositions, based on a knowledge
- Elimination of impossible propositions
- Selection and verification of the most appropriate propositions

Thus, if we want to help human intellect, to make it more powerful and more creative, we should make a knowledge machine which could assist during these steps. Let’s name demands to such a machine.

6.5. 11 demands to Knowledge Machine

Step 1 - Observation.

1. A knowledge machine should have maximum possible information about a case before a judgment.

Step 2 - Producing propositions, based on knowledge.

2. A knowledge machine should possess maximum possible knowledge in a sphere of implementation.

3. A knowledge machine should possess no excessive knowledge, should have nothing but the tools which may help in doing work.

4. Getting indication of the course of events, a knowledge machine should be able to guide itself by other similar cases which occur to its memory.

5. A knowledge machine should have an ability to take into account not only descriptions of situations in its memory but results as well, providing a possibility to reconstruct a description from a result, i.e. if you told it a result, it would be able to evolve what the steps were which led up to that result.

6. Possessing information about the great number of cases, a knowledge machine should have an ability to find a strong family resemblance about them, i.e. to find templates of typical cases.

7. A knowledge machine should have an ability to explain the grounds of its conclusion.

8. A knowledge machine should arrive at the conclusion for a few seconds after getting a description of case.

9. A knowledge machine should focus on the most unusual in descriptions of situations.

Step 3 - Elimination of impossible propositions

10. A knowledge machine should have an ability to point out all impossible propositions.

Step 4 - Selection and verification of the most appropriate propositions

11. A knowledge machine should estimate a level of a confidence of its propositions.

We think that there are many possible solutions for estimation, but we developed our own Proposition Value Index, based on idea of member of USSR Academy of Science M. N. Livanov from Russia that the essence of memory associations is a spatial-temporal coherence of narrow-band periodical oscillations of central neurons sets activity (see [3]).

6.6. AI expert systems and neural networks

Expert system, as we understand, is based on the idea of decision tree, when, with every answer to a program's question, a direction of moving through a tree changes until a final leaf (decision) will be reached (see [1]).

- So not all possible questions will be asked, and not maximum information will be received.
- The key elements are decision rules, but no knowledge itself. Not a word about the thousands of other similar cases, about typical cases.
- As we see, expert systems originally were designed to be deduction machines. But it is not very reliable to entrust to machine deciding what is absolutely impossible. We think that more fruitful approach is to show what reasons to consider some hypotheses as impossible. And only man should make the final decision.

It is not amazing that development and implementation of a successful expert system is very hard work, because experts cannot think, as a rule, in terms of decision trees, and the mathematical theory of probability have a little in common with a feeling of a confidence of an expert.

Neural network is based, as we know, on the idea of teaching of set of elements (neurons), controlling conductivity between them (see [2]). Teaching is going under control of expert, which defines whether attempt is successful. This is more merciful towards expert - nobody is trying to make him feel himself deficient asking: what is the probability of this conclusion when that parameter's value is present. But there are some difficulties, not outdone yet.

- A neural network is oriented on decision rules rather than on knowledge itself. So there are no thousands of other similar cases in memory of neural network.
- A neural network cannot explain reasons of own conclusion in terms that people can understand. So it is very hard to verify its activity and, therefore, to believe.

An expert system is an example of a 'top-down' approach when particular instances of intelligent behavior selected and an attempt to design machines that can replicate that behavior was made. A neural network is an example of 'bottom-up' approach when there is an attempt to study the biological mechanisms that underlie human intelligence and to build machines, which work on similar principles.

GKM-EKP technology is based on principles uniting both ‘top-down’ and ‘bottom-up’ approaches.

6.7. Results

The set of tools called General Knowledge Machine (GKM) was developed providing intelligent e-knowledge base engine for any kind of knowledge-based applications, supporting effective knowledge presentation, precise knowledge search, adaptive learning and immediate consulting. GKM could be used for a creation of effective knowledge-based applications called e-knowledge systems. Early versions of GKM were developed for UNIX, MS-DOS, Windows operating systems. The latest version supports all platforms of GNU compiler options (any Windows, Linux, Unix ...).

There are working products which can be presented to experts in corresponding areas.

Products were tested in various environments – business, medicine, arts. Papers were published in Russia, Italy and UK. The work was featured in the 2006-2007 Edition of the Marquis Who's Who in Science and Engineering as a pioneer research.

6.8. Conclusion

Some people say about a crisis of Artificial Intelligence. But is this crisis of human intellect? Of course, no. May be it’s a crisis of human self-confidence. In the beginning there were many promises to built machines more intelligent than people. And those machines should use advanced principles of work, much better than obsolete human intellect (see [5]). Instead of help to human intellect there were attempts to replace it. But those, who read works of academician V. Vernadsky from Ukraine ([6]), E. Le Roy ([7]) and P. Teilhard de Chardin from France ([8]), know that the main result of evolution on Earth is a creation of Noosphere - a sphere of intellect. And, in this case, it is very interesting what can be called an intellect, but is based on other principles than developed by evolution?

References

[1] J.L.Alty and M.J.Coombs. *Expert systems. Concepts and examples.* The National Computing Centre Limited, 1984.

[2] Geoffrey E. Hinton. *Learning in parallel networks.* Byte. By McGraw-Hill, Inc., New York, 1985.

[3] M. N. Livanov. *Spatial Organization of Cerebral Processes.* John Wiley & Sons: Chichester. 1977

[4] K. M. Golubev. *Adaptive learning with e-knowledge systems*. Inderscience Enterprises Limited, UK, in IJTM, Vol. 25, Nos.6/7, 2003

[5] Roger Schank, Larry Hunter. *The quest to understand thinking.* Byte. By McGraw-Hill, Inc., New York, 1985

[6] Vladimir I. Vernadsky. *The Biosphere*. tr. David B. Langmuir, ed. Mark A. S. McMenamin, New York, Copernicus, 1998

[7] E. Le Roy. *Les origines humaines et l'evolution de l'intelligence*. Paris, 1928

[8] P. Teilhard de Chardin. *La place de l'homme dans la nature*. Éditions du Seuil, Paris, 1956.

From Community Memories to Corporate Memory

Galarreta Daniel[1] and Rivière Pascale[2]

ABSTRACT

In this paper we sketch a solution to the cohabitation – and the collaboration – of two types of memory: community memories on one side and a corporate memory on the other. In practice we will consider communities of practice such as the 19 CNES Technical Competence Centres (CCT) and a corporate memory such as the one which is today managed by CNES. We will identify characteristic features that distinguish communities of practice from the overall company: size, homogeneity, temporality, dialects, tools, and ethics. We will show that the articulation of these two types of memories elucidates their necessary connection. This connection will as well offer the end-user a natural and simplified access to his/her knowledge. We will illustrate this point with a concrete case at CNES. And last, we will argument how this mutual visibility, opens promising perspectives to innovation processes.

1 INTRODUCTION

According to G. Van Heijst & al [1]: "a corporate memory is an explicit, disembodied, persistent representation of the knowledge and information in an organization". This definition deliberately restricts the form of the corporate memory since the goal pursued is to investigate how computer systems can be used to realize corporate memories. In view of this, any piece of knowledge or information that contributes to the performance of an organization could (and perhaps should) be stored in the corporate memory.

However, this approach tends to underestimate both the questions of how to use it efficiently and how to organize it to achieve its goal viz. "contribute to the learning capacity of organizations" [1]. Indeed the authors note that "one of the most difficult issues, is that there is a need for a method that structures knowledge at the ``macro level" (the level of knowledge items)"

Other definitions of corporate memories insist upon the role played by the actors, and one may sustain that the memory of a group, insofar as it includes knowledge, cannot extend outside the group of actors that interact or share a common practice.

Therefore, one may question the fact that there really exists such a thing as corporate memory, except in the form of "an explicit, disembodied, persistent representation of the knowledge and information in an organization".

We will show that not only the articulation of community memories with a corporate memory solve the issues that G. Van Heijst & al noted, but also elucidate the necessary connection of these two sorts of memory. This connection will also offer the end-user a natural and simplified access to his/her knowledge.

We will start by describing the Genealogy of community knowledge: the social frameworks of the memory (ch. II) then we will describes the Information-loaded objects for working communities (chap. III), in Chapter IV we will provide a criticism of the familiar notions of data, information knowledge; in chapter V we turn to Practical Consequences that we will illustrate with a concrete case at CNES; in chapter VI, we define corporate memories in this context. We will end this paper by final remarks and a conclusion (chap. VII). We will argument how these mutual visibilities between these two kinds of memories, opens promising perspectives.

2 GENEALOGY OF COMMUNITY KNOWLEDGE: THE SOCIAL FRAMEWORKS OF THE MEMORY

Maurice Halbwachs, observed that the groups that an individual belongs to, afford him/her the means for remembering facts or events that once happened to him/her.

But he stressed that this efficiency depends upon the fact that the individual agrees with those means and adopts at least temporarily, the way of thinking of those groups. [2]

Concerning the issue of a collective memory, M. Halbwachs went even further. It is not only sufficient to posit and observe that while remembering, individuals always use social frameworks. As far as a collective memory one must use the group as its reference. But the two aspects are closely related: one can say that an individual is remembering something by adopting the point of view of the group or conversely the memory of a group is implemented through personal memories [2]

These observations, provided we admit them, seem to be corroborated by the long term existence of groups dedicated to the transmission of knowledge and know-how, such as those that formed for instance, during the Middle-Ages, in France, the system of companionship. A prescription for the shoesmakers of Troyes in 1420 is mentioned where "several companions and workers of this trade, of several languages and nations, came and went from town

[1] Direction du Système d'Information, Centre National d'Etudes Spatial Toulouse, France, daniel.galarreta@cnes.fr

[2] KM & ECM Consulting T-SYSTEMS, Toulouse, France, pascale.riviere@t-systems.fr

[3] "Plusieurs compaignons et ouvriers du dit mestier, de plusieurs langues et nations, alloient et venoient de ville en ville ouvrer pour apprendre, congnoistre, veoir et savoir les uns des autres"

to town working to learn, know, see and update their knowledge from each other"[3] [3]

In 1991 Jean Leave and Etienne Wenger met the position of Maurice Halbwachs by contesting the fact that learning is the reception of factual knowledge and information and proposing that learning is a process of participation in communities of practice [4].

Etienne Wenger [5] asserted that "we all belong to communities of practice. At home, at work, at school, in our hobbies – we belong to several communities of practices at any given time. And the communities of practice to which we belong change over the course of our lives. In fact, communities of practice are everywhere" [5].

Despite that acceptation, "Community of practice" usually receives the restricted meaning of "a group of people who share a craft and/or a profession". It is formalized in the sense that it is dedicated to knowledge sharing, and although you cannot contrive or dictate its aliveness, it is recommended to "cultivate" it [6].

This restricted view –although valuable – of the original conception leaves outside situations where knowledge sharing is not the first aim of the group and there is no such thing as 'cultivating aliveness'.

In other words, it leaves outside "working communities", such as the communities the existence of which is justified by the tasks or missions they are assigned to, and depends on the high skills of a few experts and/or senior experts.

3 INFORMATION-LOADED OBJECTS FOR WORKING COMMUNITIES

In order to characterize the way a working community produces and maintains its community memory, we must refer again to both the arguments of J. Lave and E. Wenger on one side and M. Halbwachs on the other.

3.1 From Wenger we learn that:

1) Knowledge is a matter of competence with respect to valued enterprise – such as singing in tune, discovering scientific facts, fixing machines, writing poetry, being convivial, growing up as a boy or a girl, and so forth.

2) Knowledge is a matter of participating in the pursuit of such enterprise, that is, of active engagement in the world.

3) Meaning – our ability to experience the world and our engagement with it as meaningful – is ultimately what learning is to produce.

As a reflection of these assumptions, the primary focus of this theory is on learning as social participation. Participation here refers not to just local events of engagement or certain activities with certain people, but to a more encompassing process of being an active participant in the *practices* of social communities and constructing identities in relation to these communities".

3.2 From M. Halbwachs we learn that:

The individual identities that we assimilate to individual memory "depend upon the fact that the individual agrees with those means and adopts at least temporarily, the way of thinking of those groups". We do believe that these means are more than conceptual entities, but are materialized in the form of cultural objects – including technical objects. During their lifetime human communities produce these cultural objects which receive their value and identity through usage and interpretative processes [2]. In turn, these cultural objects by the dependences they mutually contract and the means they require to be processed, provide and/or reinforce the identity of the community that interpret and process them.

Therefore, a community can be assimilated to the cultural objects it contributes to produce, interpret and process. Of course such an assimilation is legitimate provided we extend the acceptation of cultural object to elements as various as documents, databases or social activities and interaction. [1] With this definition, memory should be better considered as a (collection of) interpretative process(es) adapted to a collection – a deposit – of cultural objects rather than an information storage technology. The shape this collection takes depends on the way these cultural objects sediment, that is to say, the history of the community itself.

The way of thinking of those groups mentioned by M. Halbwachs, is related to what one usually defines as the point of view currently admitted by those groups.

Using this simple notion of "point of view", it is possible to give a better insight of what we define as "cultural" or "information loaded" objects. We will then refer to apparently familiar notions such as *data*, *information* and *knowledge*.

4 DATA INFORMATION KNOWLEDGE: CRITICISM OF FAMILIAR NOTIONS

Depending on the working community that we consider, the "cultural" – or technical – objects may be as different as:

- A traveling-wave tube (TWT) or traveling-wave tube amplifier (TWTA) or antenna – in Radio frequency (RF) spectrum and Radio Communication domain,
- Fuse cord, explosive bolt or ignition charge – in Pyrotechnics domain.

The **descriptions** – or **views** – we can give of them can range from simple *information* to a true *piece of knowledge* and are almost always complemented with a particular kind of description that we called *data*.

We will therefore extensively discuss these notions in this chapter, since they condition the way we imagine to store, share, or preserve access to the memory – of the objects – of one working community. First of all, let us consider the ***usual distinction*** between data, information and knowledge.

Data: (from Wikipedia) can be defined "as being discrete, objective facts or observations, which are unorganized and unprocessed and therefore have no meaning or value because of lack of context and interpretation."[7]

Alternatively, (Wikpedia suggests [8]) data can be viewed either as "sensory stimuli, which we perceive through our senses"[8] or "signal readings", or as symbols. In the latter case , "symbols",[8] [8] or "sets of signs […] represent empirical stimuli or perceptions",[8] of "a property of an object, an event or of their environment"[8]; Data, in this sense, are "recorded (captured or stored) symbols", including "words (text and/or verbal), numbers, diagrams, and images (still &/or video), which are the building blocks of communication", the purpose of which "is to record activities or situations, to attempt to capture the true picture or real event.

E.g.: 37.5 ° C

Information: can be defined as an interpretation of data or their meaning in a given context. Nuances can be introduced according to the precise definition of the data that is adopted.

E.g.: the temperature of the patient is 37.5 ° C.

Knowledge: is related to the way data and information are combined in order to attain a given goal while satisfying epistemic principles such as rationality principle or a truth principle provided by experience, reasoning or even faith.

E.g.: If the temperature of the patient is above 37.5 ° C, he has fever.

Are these definitions satisfying? They are, to the extent we need to provide a meaning to the corresponding terms in order to distinguish between them. However they are not always adequate in situations where management operations such as "data preservation", "knowledge transmission" or "knowledge preservation" are needed. The reason is that the above definitions rest on intuitions or a scientific culture that are not always adapted to the practical needs we mentioned. Let us clarify this issue.

The study of the conditions of knowing, in general, constitute an important branch of philosophy, epistemology. It includes in particular, the philosophy of sciences as a sub-branch. If we restrict to that "sub-branch", it is admitted to consider that any well-formed theory starts from raw material – its object – that constitutes its first level. Then this theory produces a description of this material. That description constitutes the second level of the theory. Its third level defines its descriptive concepts and constitutes the methodology level of the theory. Its last level is the epistemological level. On this level, the soundness of the methodology is criticized and verified by testing its coherence and by measuring its adequacy with respect to the description, moreover and among other things; description discovery procedures must be evaluated [9]

Clearly the notions of data and information correspond to the epistemological level (and not to the methodology since "data" and "information" do not correspond to descriptive concepts – except in information theory). For instance, concerning data (as facts), the article already quoted from Wikipedia specifies: "Insofar as facts have as a fundamental property that they are true, have objective reality, or otherwise can be verified, such definitions would preclude false, meaningless, and nonsensical data from the DIKW" This formulation is clearly an epistemological statement concerning data, and moreover, corresponding to a positive epistemology.

Since Information is concerned by the condition of interpretation of data, it is also related to the epistemological level.

Knowledge, as it was defined above, is almost a rephrasing of the definition of the epistemological level itself.

In short, the definition of data, information and knowledge, corresponds to the definition of an epistemology, but let us underline that it does not correspond to any possible epistemology. We noted for instance that data – as facts – correspond to a positive epistemology, namely a conception of science that is well suited for classical physics. However, we know that within other scientific domains such a conception is not relevant: for instance for quantum physics, or in the study of living organisms, or in the domain of social and human sciences.

It is not the intention of this paper to further develop this issue, but it is now easier to admit that these usual definitions of data, information and knowledge we recalled, are not really suited for practical knowledge management even if they sound so. The reasons have been already given: as long as we situate ourselves within the domain of classical sciences, there are no objections to these definitions. As soon as we turn to social or human sciences or practices these definitions are not useful except for their rhetorical virtue to convince that the three notions should be distinguished.

For instance there is no serious reason to sustain that the complete work of Balzac, the Human Comedy – La Comédie humaine – corresponds more to information than to data or to knowledge.

The "Comédie humaine" consists of 91 finished works (stories, novels or analytical essays) and 46 unfinished works (some of which exist only as titles) (source wikipedia [10]): Can be viewed as data for a thematic analysis of Balzac works when supported by computer. For instance in [11] Th. Mézaille, using the program developed by E. Brunet, Hyperbase, to analyse the meaning of joy in Balzac's works. Can be viewed as information about French society in the period of the Restoration and the July Monarchy (1815–1848) ([10]): Viewed as a knowledge about speculation and finance in the 19th century in a thesis such as "A Fever of speculation, narrating finance in the Nineteenth-century novel" thesis by Tara McGann[12].

Even in the case of physical entities, it would be difficult to sustain that a piece of data such as: "the temperature is 37.5 °C", does not involve any knowledge about what the "temperature" is. This knowledge necessarily includes the fact "°C" meaning "degrees Celsius", that 37.5° C is above 0°C which is defined as the freezing point of water but is under 100 °C which is defined as the boiling point of water. This knowledge may extend to the fact that these bounds are defined with respect to the atmospheric pressure at mean sea level. Therefore this temperature could not characterize elements belonging to cryogenics, nor to solar physics.

All the above analyses and criticisms, hinge upon the fact that the distinction between data information and knowledge depends upon the viewpoints that are applied to what Peter Stockinger [13] calls information-loaded objects, or that can be also called "signifying objects":

An arbitrary and limited list of examples of such information-loaded objects could be:

- Written documents (memos, meeting minutes, requirements documents, reports, contracts, norms, text-books, ...)
- 2-D, 3-D formalized representations (organograms, work breakdown structures, Gantt and Pert modelling of a planning, mind map, UML representations, electric schemes, architecture plans, 3D virtual models of a physical equipment, …)
- Images (analogic signal representations, pictures, drawing, satellite images, …)
- Video images
- Sound recordings
- Oral discourses
- Bodies in nonverbal communication (individual and collective manners to use one's body cf. [14])
- Tools and/or Interfaces of those tools (machine tools, vehicles, …)
- Architectural implementation of a building
- Urban organization of a plant, of a city

Therefore, the distinction between data, information and knowledge must not be attached to the nature of these information-loaded objects. The distinction should be more adequately based upon the interactions of the viewpoints upon these objects as simply as that:

- Whenever there is a confrontation between viewpoints, a view about an information-loaded object corresponds to a **piece of information**
- Whenever this confrontation evolves toward a negotiated agreement, a view about the object corresponds to **a piece of knowledge**
- **A piece of data** corresponds to a view of the object from just one viewpoint when other possible interactions are suspended. - (Precisions about these questions can be found, for instance in [15]).

One could wonder about the existence of an object considered as information-loaded objects – signifying objects – outside the existence of views (of this object) in the form of *pieces of information*, *knowledge*, or *data*. Such an issue will not be developed here ([15] for this). However from the position we adopt here, we want to identify the views this object with the object.

More precisely, as far as a minimum of shared visibility of this object between their potential users is required in order for those users to agree about a common identity, then we admit that a minimum of agreed views – about this object – must be formed by the users.In the following, we designate these agreed views as a *knowledge object*.

5 PRACTICAL CONSEQUENCES

Let us consider the case of *the working community* that has been identified during the knowledge transfer between a senior expert (J. Sombrin) in the Radiofrequencies domain and other colleagues (often young) belonging to the department of Radiofrequencies at CNES – the French space agency. The phrase *working community* receives here the acceptation we gave above, namely as "*the communities the existence of which is justified by the tasks or missions they assigned to, and depends on the high skills of a few experts and/or senior experts"*.

This knowledge transfer was justified by the retirement of the expert. It had been identified that most of his expertise was crucial for the department and colleagues, specialists in that domain. A mapping of his knowledge was performed: a survey of the information, knowledge and competencies held by the senior was done, then an assessment of their criticality with respect to the tasks to be performed and the fact that no redundancy was available. This mapping was done concurrently both by the senior expert and the specialists. During this process a *common view* of the organization of the knowledge of the domain from the points of view of processes, content, and technology.

By doing that, mapping reveals a *working community* that *potentially exists*. It actualizes – concretizes – a community that we can suppose rests upon oral practices of communication. By doing that, mapping offers an "information-loaded object", namely a ***map***, that results from the *negotiations* between the different viewpoints involved – the senior expert and his colleagues. This map corresponds for each point of view to a visible *knowledge object*.

Thanks to that *knowledge object*, the *map*, the organization – the CNES, in the considered example –, then becomes aware of the necessity of building a tangible representation of *a working community* which will correspond to the existing one. Ideally this representation should enable the existing community to share the same technical objects as well as the same knowledge – and know-how – and perform the same usual tasks as well as the tasks to come in the future.

The current natural tool of the promotion of the information is the Internet/intranet technology. Using the available tools in CNES in order to supply the community with such are-presentation seems rather natural. It was then suggested – in our Radiofrequencies case – to use the collaborative environment provided by the Livelink solution (an Open Text product) which is available to us.

Insofar as we adopt the view that a "community can be assimilated to the cultural object it contributes to produce, interpret and process" (see above) two lines of solution are available to us – and may be combined –:

- either we make already existing cultural or knowledge objects visible for an existing and corresponding community
- or we propose new cultural or knowledge objects that can be identified as acceptable translations or transpositions of original existing objects, and satisfy a visibility condition.

In any case, a community must be able to self-identify itself thanks to those knowledge objects. Besides that, other existing communities must be able to identify that community thanks to few or many shared knowledge objects. Such existing communities may be for instance:

- Centres of technical competencies – Communities of Practices that were created by CNES
- Research communities or organizations such the Scientific National Research Centre (CNRS)
- Corresponding communities in industry

The sharing of knowledge objects between communities is limited by the understanding that one community can have of the knowledge objects of another community. As we pointed out above, the knowledge objects depend on the interactions that the different points of view may have all together. Therefore, the more one community have interactions with another community, the higher the probability for both communities to share knowledge object.

Conversely, as soon as we turn to communities that have few contacts with one of the preceding communities, the sharing of common knowledge objects tend to rarefy.

The interactions of points of view that produce knowledge objects are not restricted to inter-community interactions. They also concern intra-community interactions. If these objects of knowledge form quite a coherent and understandable collection for the community at a given time and if they are highly dependant upon that time, their sustainability depends strongly on the continuity of the community in time. A significant change within this community for various reasons (change of place, organization, people) can cause the loss of the understanding of these objects of knowledge.

In the definition we adopt for a working community, we underline the fact that its existence *depends on the high skills of a few experts and/or senior experts*. In the case we considered that community hinge on J. Sombrin an internationally renowned expert in relation with other experts in the Radiofrequencies domain. In other words the members of such a community need to recognize their peers – and may be leading ones. If the number of people increases it is questionable if the notion of "working community" is still applicable in this new situation. Even if it is difficult to assign a limit to the size that such a community should not exceed, in order to deserve its original qualification, such a limit exists.

Does any *community* exist beyond that limit – the value of which may depend upon the domain? Very likely the answer is yes.

Let us remember that the general definition of a community we adopt is as above: *a community can be assimilated to the cultural* *objects it contributes to produce, interpret and process.* The increase of the size has a direct impact of the *range of knowledge objects* that the individuals are forced to refer to in professional transactions in order to carry out their project. That is to say an impact on their *strategies of communication*.

We admit here that the scientific and technical ***concepts*** that experts normally use in the course of their activity, do not depend on the size of the concerned community: strictly speaking their specialised language (or language for specific purpose – LSP), does not depend on that size. It is the way that this community is organized and names the information it stores that will change as the size of the community is increasing. In a proper working community, this way is mostly influenced by the relations the experts have between them where oral modalities of communication (i.e. orality) plays an important role in organising and naming the categories of storing. More precisely, orality tends to hypostatize categories that are simply accidental, or contingent. The cause is that simple orality, is highly dependent on contingent situations of narrative character and this modality is efficient provided there exists oral repetitions of the admitted categories.

When the size of the community increases, it is more difficult to have common histories among its members and to adopt and share categories that organize and name the information the community uses. In theses new conditions, scientific and technological concepts as well as the specialised language that the experts share, become the principal reference between the experts whatever the size of the community.

It is that situation that justifies use of a corporate memory.

6 CORPORATE MEMORIES

Let us consider again the knowledge objects. They are not necessarily concrete objects but rather *places* – in the sense that this word receives in the term *common places* – where different points of view find a minimal agreement by producing mutually agreed views. From this minimal agreement – or places – can stem more concrete objects.

The longevity – and stability – of an *knowledge objects* within a community is directly related to the number of views and points of view that produce them, and to the extension of the period of time during which those views are produced.

In this perspective, the collection of texts that a community produces during its life time is the best material that points of view can exploit in order to provide numerous views, but in the same time it is rather difficult to process extensively. Texts written in natural language when considered for their own sake, independently of a precise and limited context of use, need specialized competencies to be fully exploited, namely in linguistics and knowledge engineering. These skills are combined in order to produce terminological models of technical domain (e.g. Radiofrequencies) and correspond to activities of interpretation of the collection of texts available for that purpose.

What precedes justifies the existence of a *socio-technical facility* that we designate as a *corporate memory* that possesses at least these features:

- A basic function: the conservancy of knowledge objects of the company
- An associated activity: analyses of linguistic resources and modelling of terminologies
- A material: the collection of texts that the company choose to preserve

In short a corporate memory plays the role of a natural conservatory regarding the written production of the company.

More precisely we define a *corporate memory* as:

- a collection – mainly of textual character – that is devoted to the conservancy of knowledge objects that are produced by the company essentially in written form
- the activity that analyses and models terminological and linguistic resources
- the formal terminological data that results from the preceding activity in the form of ontologies, taxonomies, thesaurus, folksonomies, and so on, and are used by search or categorizer engines.
- the search or categorizer engines that use the preceding formal data.

Regarding the issues either of distance between two working communities that lack interactions or of distance between people that belong to a community that does not allowed direct contact because of its size and organization, a corporate memory facility brings a solution. By its conception a corporate memory provide an upper level that more persistent and sharable knowledge objects. That collection of objects offers a bridge above the possible brick walls between communities. It can also maintain at least temporarily a common reference between people immerged within a large community.

On the other hand regarding the issue of maintaining a coherent and understandable collection of knowledge objects within a working community in time, the corporate memory facility is naturally suited to provide such a solution. The reason is clear: it is concerned by objects that belongs to an *a priori* larger temporal scale than the ones used by usual working communities, then it can solve this problem. In the case of a working community, the time scale is approximately a decade whereas in the case of a corporate memory it could be as long as half a century – the timescale of History.

7 FINAL REMARKS AND CONCLUSION

Let us make this final observation. In all wethat presented above, the corporate memory facility plays the role of a “bus of knowledge” between working communities, providing them with ***views*** of knowledge objects it preserves.

But, we can imagine that this facility would provide the working communities with the ***knowledge objects*** themselves, in order to be adapted to those communities. Instead of producing and preserving the knowledge objects through a totally idiosyncratic process, we suggest that the knowledge objects of the corporate memory could be used as strains from which knowledge objects of those communities can be produced and maintained. Those communities could then use such objects in their search or categorize engines.

7.1 A facility for innovation

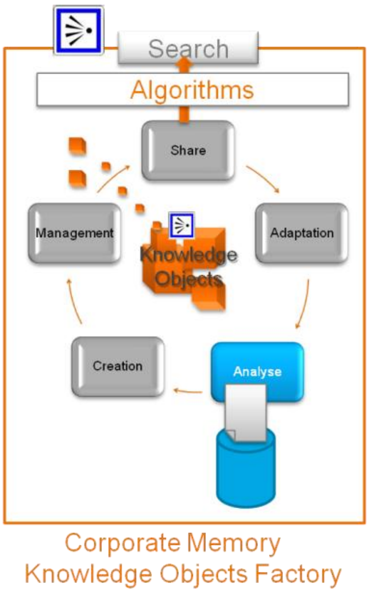
The corporate memory could also be considered as a knowledge objects factory (fig 1) with the capability to provide facilities for all communities such as:

- Tools and methodologies for analysis the collection of text, testimony of expert, …
- Processes and methodologies to create and manage Knowledge objects,
- Knowledge objects like map, terminology, ontology, modeling process …

- Formal terminological data (thesaurus, folksonomies ,…) for creation or/and optimization of algorithms to improve the Search and categorizer engines.
- Views (questions/answers) to offer at the end-users the result of their Questions/Answers.

Those objects are not static and will be adapt according to the communities, time, distance and size in a common language for efficient sharing.

Figure 1: The corporate memory: a knowledge objects factory



The corporate memory provides the different communities not only with views – for search and categorizing purposes –but also knowledge objects that these communities will be able to use, confront and combine, in order to produce ***new*** knowledge objects. The figure below (fig. 2) shows how the corporate memory:

- provides the different communities with relevant knowledge objects;
- collects knowledge objects from these communities and from external communities;
- adapts, transforms, combines these knowledge objects, in order to produce and share new and innovative knowledge objects –

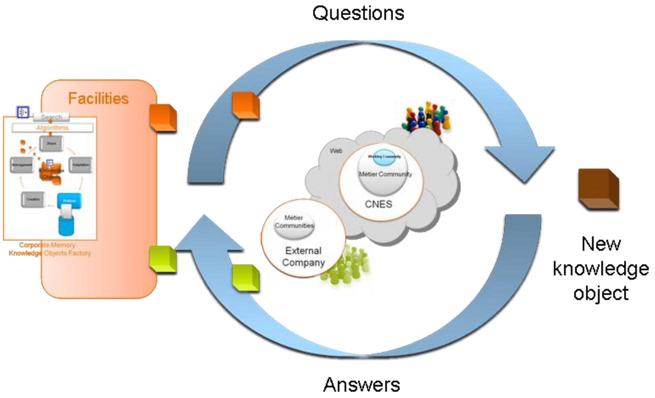


Figure 2: The corporate memory as an innovation engine

7.2 Conclusion

The main contribution in our proposal is to consider a corporate memory as a facility that naturally complements community memories that exist within a company such as the CNES. What we underlined is that they all include activities of their members – in community memories as well as in corporate memory – since the production of knowledge objects necessarily implies the participation of several points of view. The visible character of these objects in both, cases reminds us of the definition that M. Grunstein gave of capitalizing knowledge whose goal is to “locate and make visible knowledge produced by the company, be able to keep, access and update them, know how to distribute, better use and create synergy between them, and valorise them”.[1995]. This definition perfectly applies to the finality of a corporate memory, but it ignores the material that makes this possible – texts – and the specific activities that allows this finality – analyses of linguistic resources and modelling of terminologies. Regarding the view of G. Van Heijst & al [1], we insist on the fact that if a corporate memory is to “contribute to the learning capacity of organizations” [1], it should involve itself in the activity of learning the knowledge objects that these organizations produce and use for that leaning purpose.

REFERENCES

[1] G. Van Heijst, R. van der Spek, E. Kruizinga, “Organizing Corporate Memories” In B. Gaines, M.Mussen , Proceeding of the 10th Baff Knowledge. Acquisition for Knowledge-Based Systems Workshop (KAW ’96). Banff, Canada

[2] M. Halbwachs, “Les cadres sociaux de la mémoire”, Albin Michel, 1994. pp.VI-VIII

[3] F. Icher, « La France des compagnons », Éditions La Martinière, 1994

[4] J. Lave, E Wenger, “Situated Learning: Legitimate Peripheral Participation”. Cambridge: Cambridge University Press, 1991

[5] E. Wenger, “Communities of Practice: Learning, Meaning, and Identity”. Cambridge: Cambridge University Press, 1998.

[6] E Wenger, R. McDermott; W. M Snyder, “Cultivating Communities of Practice” Harvard Business Press, 2002.

[7] J. Rowley, H. Richard , “Organizing Knowledge: An Introduction to Managing Access to Information”, .Ashgate Publishing, Ltd. 2006. pp. 5–6.

[8] “DIKW”. Wikipedia, The Free Encyclopedia. Wikimedia foundation, Inc. 05 may 2012. Web 28 may 2012. http://en;wikipedia.org/wiki/DIKW

[9] A.J. Greimas, J. Courtès, “Sémiotique. Dictionnaire raisonné de la théorie du langage », Hachette, 1979.

[10] “LaComédiehumaine”. Wikipedia, The Free Encyclopedia. Wikimedia foundation, Inc. 13 May 2012. Web 29 may 2012.

[11] Th. Mézaille, « Etudier les textes littéraires numériques. introduction à une pédagogie. Enjeux et objectifs de l'analyse textuelle ». Chapitre 1 : juin 2001 [en ligne]. Disponible sur : www.revuetexto.net/Reperes/Themes/Mezaille_Etudier

[12] T. McGann, A Fever of Speculation: Narrating Finance in the Nineteenth-century NovelColumbia University, 2006.

[13] P. Stockinger, “e-semiotics”, Oy Fountain Park Ltd, Helsinki, November 23rd , 2001.

[14] M. Mauss, « Les Techniques du corps », *Journal de Psychologie* 32 (3-4). Reprinted in Mauss, *Sociologie et anthropologie*, 1936, Paris: PUF.

[15] D. Galarreta, “E-science for humanities – a semiotic framework.”, Problems and Possibilities of Computational Humanities 4-6 July 2011, Groningen, The Netherlands.

Contextual knowledge handled by an expert system

Janina A. Jakubczyc and **Mieczysław L.Owoc**[1]

Abstract. The evolution of expert systems (or more generally intelligent systems) exhibits the increasing role of context. From two main usages of context: (1) to internally manage knowledge chunks and (2) to manage the communication with the user, we concentrate on the first one. Our proposition is two folded: the first one faces the problem of automatically delivering relevant contextual and context dependent knowledge for given problem applying the contextual classifier. The second one concerns the new role of intelligent systems supporting the more active user participation in problem solving process. This proposition gives the possibilities: to more complete utilization of the knowledge closed in data bases, to solve the difficult problems and to exploit user knowledge about the problem under consideration.

1 INTRODUCTION

There are almost common opinions that potential weakness of intelligent systems consists in neglecting of multidimensional aspects of reality where these systems are implemented. In other words contextual dimensions of applied knowledge should be included in order to improve effectiveness of the discussed systems. This approach allows for more holistic view on intelligent system concepts embracing users, performed tasks, different situations creating complex environment with a crucial role of context.

There are many surveys and papers stressing increasing importance of context, for example: Brezillon[2], Cram and Sayers[3]. The list of application areas where research on context are very important refers to word recognition (as a primary and natural fields of investigation) up to contemporary economy. Some of authors stress an assistant role of such systems: Boy and Gruber[4] or Brezillon[5]. Especially such approach seems to be very useful in human-machine systems where knowledge can be acquired from users.

Our proposition is two folded: the first one faces the problem of automatically delivering relevant contextual and context dependent knowledge for given problem solving by using the contextual classifier. This approach admits of incomplete context and context dependent knowledge. The second one concerns the new role of intelligent systems supporting decision makers. It means more active participation of ES in identifying contextual knowledge relevant for users making decisions.

The paper is managed as follows. The next part is devoted to automatic acquisition of contextual and context dependent knowledge problem. Assumptions, main goals and limitations of this acquisition process are considered. A concept of contextual knowledge base in the prepared ES prototype is presented in the next part. The changing role of an user in decision making process supported by the elaborated system is presented in the next session. Then some examples of supporting selected contextual classification activities useful in the decision making process with the idea of KBS system are demonstrated. The paper ends discussion about issues arising from the research.

2 THE ACQUISITION OF CONTEXTUAL AND CONTEXT INDEPENDENT KNOWLEDGE

The knowledge acquisition for knowledge base system is still a challenge for knowledge engineers. Our proposition faces two issues. The first one is abandon the problem of direct knowledge acquisition from experts and move the focus on the automatic knowledge discovery from more and more numerous data bases that contain considered by the experts decision cases. The second one is to release the degree of acquired knowledge generalization. This will be achieved by introduction contextual dimensions to knowledge. It gives the possibility to use more correctly the knowledge in ES[6] and to structure of the knowledge base. The knowledge base enrichment of contexts gives the opportunity to solve difficult socio-economic problem for which finding acceptable single model classification model is almost impossible see Brézillon[7].

Our solution to above issues is contextual classifier algorithm. The main idea is to discover the knowledge useful in problem solving according to its possible contexts. Thus the focus is not on knowledge generalization for a class of similar problems but on its generalization in specific context. Hence the main issue is context definition and context identification.

There are many definitions of context (for detail consideration see for example Akman[8]) and one can choose according to his

[1] Artificial Intelligence Systems Department, Wroclaw University of Economics, email {janina.jakubczyc,mieczyslaw.owoc}@ue.wroc.pl

[2] Brezillon, P. (2003): Context-Based Intelligent Assistant Systems: A discussion based on the Analysis of Two Projects. Proceedings of the 36th Hawaii International Conference on System Sciences - 2003

[3] Cram J., Sayers R. (2001): Creating and Managing Context: The Use of Knowledge Management Principles to Deliver Virtual Information Services to Schools. ASLA XVII Conference, Queensland 2001

[4] Boy G., Gruber T.R. (1990): Intelligent Assistant Systems: Support for Integrated Human-Machine Systems. AAAI Spring Symposium on *Knowledge-Based Human-Computer Communication*, March 1990, Stanford University.

[5] Brézillon, P. (2011) "From expert systems to context-based intelligent assistant systems: a testimony". The Knowledge Engineering Review, 26(1) , pp. 19-24.

[6] Ibidem

[7] Brezillon P., Some characteristics of context, Lecture Notes in Computer Science, 2006, Vol. 4031, pp. 146-154.

8 Akman V., Rethinking context as a social construct, „Journal of Pragmatics" 2000, Vol. 32, No. 6, pp. 743-759.

needs. For our use, we quote after Brezillon and Pomerol[9] 'the context is that which constrains something without intervening in it explicitly'. So we can say that context is all that forces understanding and interpretation of given concept. The complexity of the context may be different from single concept to complex description.

The context identification is dependent on the context localization: internal in the data base or external referred to the domain experts and analytics (for details about the context identification see Jakubczyc [10]). If the context is internal - proposed algorithm is able to cope with its identification.

The given concept can be seen in one or more contexts thus there is the possibility to employ some mechanism to select appropriate chunks of context dependent knowledge and to combine it into the final decision. Thus we need the contextual classifier ensemble (the detailed description of contextual classifier ensemble can be find in Jakubczyc's work[11]).

The criterion for knowledge discovery is one or more contexts. It gives the possibility to utilization of more information that is included in data sets or outside data set (additional information). The number of contexts in which some problem can be perceived is finite, so it results in known number of base classifiers. The different context distinguishes base classifiers and gives the interpretability of each single base classifier and classification results.

The acquisition of contextual and context dependent knowledge algorithm consists in:

a) build decision tree on the basis of the entire learning set (basic attributes i.e. problem descriptors in the model and irrelevant attributes i.e. insignificant to the problem solving, they remain outside the model),
b) context identification: create decision tree for each decision attribute, taking into consideration only irrelevant attributes (context-sensitive attributes from basic attributes; context attributes from irrelevant attributes),
c) context qualification: identify pairs 'contextual /context-sensitive' of attributes that can be used to the partition the learning set (according to the assumed level of classification accuracy),
d) build contextual base classifiers for each selected context as the compound of decision trees generated for each learning subset of the selected context.

As the results we obtain chunks of knowledge that describe identified context and chunks of context dependent knowledge. The former include knowledge chunks of contextual situations and the latter include context dependent knowledge chunks of specific contextual situation.

Taking into account the lack of consensus how to select and combine contexts (the criterion of diversity, the criterion of the number of context dependent knowledge chunks and their classifier accuracy) number of possible contextual classifier ensembles with a similar level of classification accuracy may be too large to decide about arbitrarily choosing one or several of them. In this we see the role of an user in problem solving process. In order to do this structure organization of knowledge base should be prepared.

3 KNOWLEDGE BASE STRUCTURE

The knowledge base consists of identified context models in the form of decision trees (Fig.1). As we see, the contexts may have different degree of complexity. The context determines set of contextual situations each described by the tree path from the root to the leaf (using the decision tree terminology).

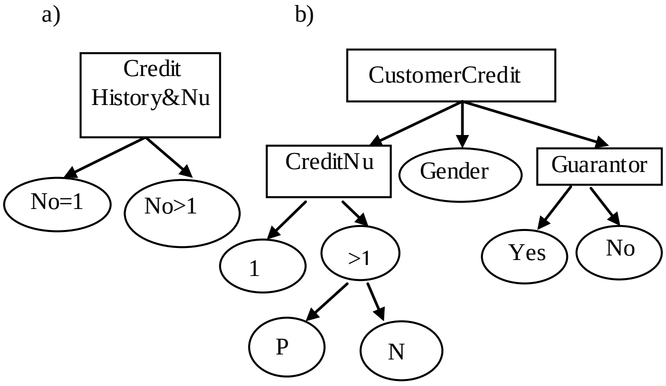


Figure 1. Examples of context structure

When the context describes one concept, the number of contextual situations is determined by the set of concept's value. In the case shown in Figure 1a), two contextual situations form two context dependent knowledge chunks. The context 1a) determines the relation between credit history and the number of credit taken. It divides the credit clients into two groups. The first group consists of people with good credit history that have one credit at the most. The second group would have trouble with credit history because of number of credit taken is higher than 1. Each of contextual situations can generated different quality context dependent knowledge chunks.

In the case of the more complex context (see Figure 1b), each path from the root to leaf determines the contextual situations and the partition of context dependent knowledge chunks (the number of contextual situation is 7). The context 1b) describes the customer participation in the others credits as co-applicant or guarantor.

We may say that each context determines the contextual situation and each contextual situation determines chunk of context dependent knowledge (Fig.2).

Introducing the contexts to knowledge base implies the partition of knowledge base and more detailed view on problem solutions. The knowledge base build on the basis of contextual classifier differs from that acquired from domain experts by a knowledge engineer. Let's consider the following issues:

– the changeability of knowledge base,
– the quality of knowledge base,
– the structure of knowledge base.

The potential resources for discover the context and context dependent knowledge are data bases that change from time to time. The frequency of such changes differs across the knowledge domains. Thus starting over contextual classifier algorithm for knowledge discovery with the suitable frequency is the necessity.

Hirst G., Context as a spurious concept, [in:] Paper Presented at the AAAI-97 Fall Symposium on Context in Knowledge Representation and Natural Language, MIT, Cambridge, Mass., 1997.

[9] Brézillon P., Pomerol J.-Ch., *Contextual knowledge sharing and cooperation in intelligent assistant systems*, „Le Travail Humain" 1999, Vol. 62, No. 3, pp. 223-246.

[10] Jakubczyc J.A., The context in concept learning, [in:] Nycz M, Owoc M.L., (eds.) *Knowledge Acquisition and Management*, Research Papers of Wrocław University of Economics No. 133, Publishing House of the Wrocław University of Economics, Wrocław 2010, pp.42-57

[11] Jakubczyc J., *Contextual classifier ensembles*, [in :] Abramowicz W. (ed.), *LNCS 4439 – Business Information Systems*, Springer, Berlin, Heidelberg 2007, pp. 562-569.

There is no increasing knowledge base modification as in typical ES, there change all.

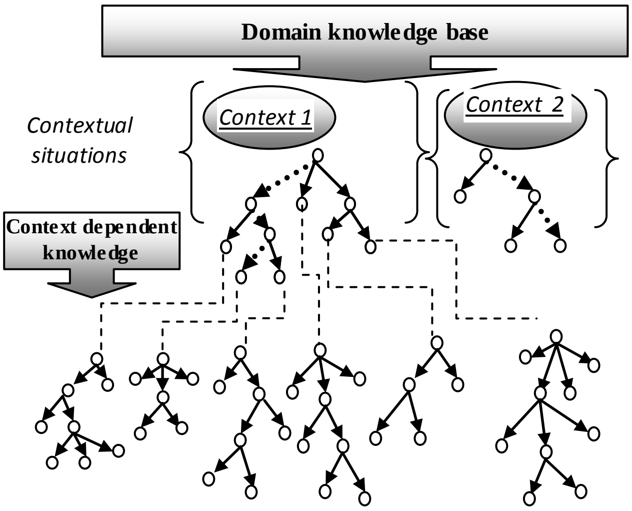


Figure 2. The structure of knowledge base

The potential resources for discover the context and context dependent knowledge are data bases that change from time to time. The frequency of such changes differs across the knowledge domains. Thus starting over contextual classifier algorithm for knowledge discovery with the suitable frequency is the necessity. There is no increasing knowledge base modification as in typical ES, there change all.

One can't say that employed data bases are representative sample of the population in statistical terminology because of different purpose (intended use). Therefore the quality of discovered knowledge may differ significantly. The measure for knowledge piece quality is based on the classification accuracy since the knowledge chunks are represented by the decision trees models. Each chunk of knowledge (context and context dependent) has assigned that measure. There are knowledge pieces that have the same or much the same measure value, there are knowledge chunk that have unacceptable measure value, as follows from our research. The level of knowledge quality determines the way the knowledge is handled in problem solving process. If the quality of single context dependent knowledge is over 80% classification accuracy user can select such chunk of knowledge as solution. More frequently we encounter the cases when the level of quality is merely above chance i.e. a little more than 50% classification accuracy. Then there the problem solution has to be created as an ensemble of knowledge chunks. Such approach gives the possibility to reach the acceptable level of solution quality.

The structure of knowledge base is determined by the type of context. As we mentioned earlier the relations between identified contexts may differ from heterogeneous set of context to the sets of contexts that are different by their granularity. The latter context relation can't be discovered by proposed algorithms, thus the domain expert or analyst may be resource of acquiring knowledge about this type of context.

Differently than in typical knowledge base there can be many relevant pieces of knowledge for one problem situation. Thus to find solution to the problem under consideration one have to select one or more appropriate contexts to view the possible knowledge pieces referring to problem solution. For these tasks we employ the user.

4 THE USER PARTICIPATION IN PROBLEM SOLVING

The user participation in problem solving process concerns three possible versions (version I, version II and version III) with different level of required activity. The contexts possible strength of influence on problem solving, quality of generated knowledge for each context and more detail for each contextual situation distinguish these versions. The first takes place when there are none identified contexts or user was searching for solution in the context but resign on behalf of not contextual solution. In this case user applies domain knowledge that was generated for entire data base if the level of classifier accuracy is acceptable. The role of user thus is limited to introduction of the description of decision situation.

The second version takes place when all context dependent knowledge chunks generated on the basis of identified contexts have acceptable quality for single or ensemble solution building. This means that each context constitutes a consistent whole. The user participation in this case embrace analysis of existing context, choice one or the set of the most adequate contexts, assessment of created the possible ensembles for chosen contexts, selection a solution of the problem under consideration.

In the real world above situations not have to be a rule, so there is a need for one more version. The third version deals with more difficult case i.e. when only some of context dependent knowledge chunks reach acceptable level of accuracy. This means that there can be context for example with twelve contextual situations from which only four determine context dependent knowledge chunks with acceptable quality level. This case is very demanding for the user i.e. the user is to conduct more profound analysis of contexts and contextual situation, choose more detailed contextual situation instead of whole contexts, assessment more possible ensembles and choice appropriate solution.

To make the user participation the most comfortable we propose two algorithms to support him in the problem solution process to handle version II and III (the version I is automatic). The algorithm for version II and is rather simple and consists of the following steps:

1. *Introducing description of the problem*[12]:
   a. phase I – search for adequate contexts according to problem description
      - presentation of contexts in which problem solution may be perceived in the form of list with the context's name and its descriptors and complexity (number of contextual situations)
      - if needed the selected context may be presented as decision trees or set of rules for more detailed view
      - *choosing single context or set of contexts for creating problem solution*
   b. phase II – presenting solution for single context or creating possible ensembles for problem solving.
2. *Selection of fusion schemata*
3. Listing the ensemble in accuracy and complexity order
4. *The evaluation of solutions due to the relevance of the classification accuracy, the complexity and understandability.*
5. *Accepting single solution or choice the best of classifier ensemble.*
6. Presentation the paths of the chosen problem solution.

[12] The user tasks are written in italics

Algorithm for user supporting in process solving for the variant III embraces:

1. *Introducing description of the problem*
    a. phase I – searching for adequate contextual situations according to problem description
        - presentation of contexts in which may be perceived problem with listing acceptable contextual situations
        - if needed more detailed presentation of chosen contextual situations and corresponding knowledge chunks in the form of listing name of context and its descriptors and knowledge chunk in the form of decision tree or list of rules
        - *choice of the most adequate set of contextual situations*

    b. phase II – creating possible ensembles for problem solving.
2. *Selection of fusion schemata.*
3. Listing the ensembles in accuracy and complexity order
4. *Evaluation of the effectiveness and expectations of received solutions.*
5. *Verification and final acceptance or return to the beginning.*
6. Presentation the paths of the chosen problem solution.

As we can see, the user has to be familiar with domain knowledge of solving problem for variant II and III. The variant I do not has such requirements about the user competency. But it seems that even for more automatic knowledge systems domain specialist are welcome.

We assume supporting essential for an user contextual knowledge by elaborated expert system where domain knowledge exists along to pieces of contextual knowledge bases.

The tool that meets the requirements specified in the form of algorithms defined in the next section is SPHINX where expert system application can be elaborated. SPHINX is the domain-independent tool for building ES[13]. It is based on blackboard type of the approach and has the ability to combine many different pieces of knowledge and the ability for consistent reasoning. The platform provides control of backward reasoning, what in the case of the contextual classifier allows someone to simulate and verify the different variants of the decision-making process.

The following features of the choice of this system, as a tool for the implementation of the proposed approach, are crucial:

• array architecture,
• easiness of applying parametric variables,
• ability to define hybrid systems.

From the structural point of view CKMES (Contextual Knowledge Management Expert System) two main categories are typical in this application:

a) domain knowledge base,
b) one or more contextual knowledge bases,

Therefore typical functions of such application embraces:

- inserting formalized domain knowledge
- incorporating one or more contextual knowledgebases
- analyzing content and relationships between components of contextual knowledge introduced to the system
- cooperation between an user and ES application in generating contextual-dependent decisions

Both types of KB domain KB as well as contextual knowledge should be located applying mentioned blackboard approach. Structure of such concept is demonstrated in the Fig.3 as knowledge sources.

Each context is described and represented by a single source of knowledge, saved in a separate file, named adequately to the context name (Fig. 3).

```
Knowledgebase of BankCustomers
Sources:
        Context1:
                type kb
                file"b11_e.zw"
        Context2:
                type kb
                file"b21_n.zw"
        Context3:
                type kb
                file"b27_m.zw"
        Context4:
                type kb
                file"b30_i.zw"
        Context5:
                type kb
                file"b25_c.zw"
        Context6:
                type kb
                file"b16_e.zw"
end;
```

Figure 3. Contexts as external sources of knowledge base

Preliminary work on the implementation of these algorithms is given in section 5.

5 THE EXPERIMENT

The idea of proposed approach is presented on the exemplary problem of credit scoring on the basis of the German Credit data[14.] The data contains 1000 past credit applicants, described by 30 variables. Each applicant is rated as “good” or “bad” credit in the ‘decision’ variable. The variables were decoded from their original names (A91, A92, A93 etc) to reflect the values they actua lly represented for more legibility.

The idea of proposed approach is presented on the exemplary problem of credit scoring on the basis of the German Credit data[15.] The data contains 1000 past credit applicants, described by 30 variables. Each applicant is rated as “good” or “bad” credit in the ‘decision’ variable. The variables were decoded from their original names (A91, A92, A93 etc) to reflect the values they actually represented for more legibility.

5.1 Knowledge acquisition

The first step is the preparation data base for expert system. Thus there is contexts identification, thus there decision tree model for all data is build. The knowledge in the form of decision tree form of set of rules for the system is depicted respectively on fig. 4 and fig.5. The decision tree presented in fig. 4 shows nodes crucial in decision making processes and relationships between them and in

[13] http://aitech.pl/content/view/48/36/lang,ISO-8859-2, 01-12-2009

[14] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[15] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

such a way the general concept of discovering knowledge is visualized. The customer evaluation depends on six attributes, i.e. checking account, credit history, other debtors or guarantors (coapp), duration (time in this bank), savings, property (the kind of credit secure).

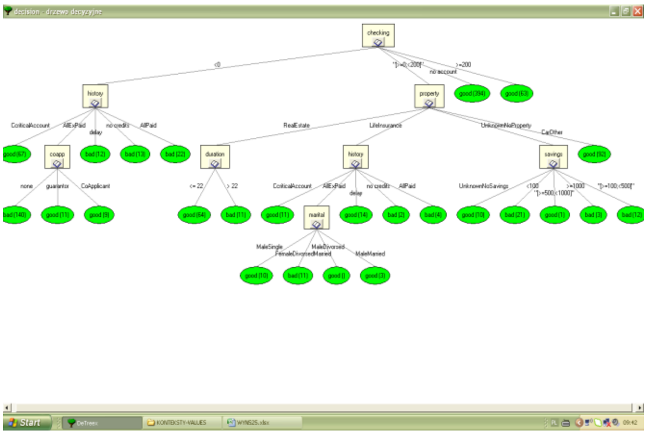
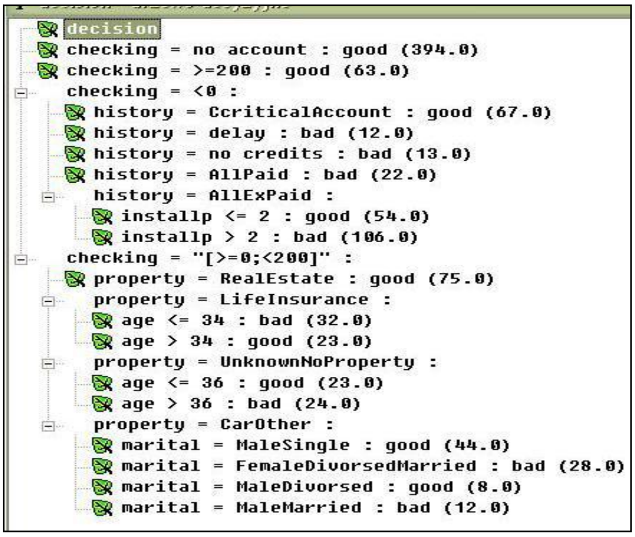


Figure 2. The decision tree for credit scoring



```
decision
checking = no account : good (394.0)
checking = >=200 : good (63.0)
checking = <0 :
    history = CcriticalAccount : good (67.0)
    history = delay : bad (12.0)
    history = no credits : bad (13.0)
    history = AllPaid : bad (22.0)
    history = AllExPaid :
        installp <= 2 : good (54.0)
        installp > 2 : bad (106.0)
checking = "[>=0;<200]" :
    property = RealEstate : good (75.0)
    property = LifeInsurance :
        age <= 34 : bad (32.0)
        age > 34 : good (23.0)
    property = UnknownNoProperty :
        age <= 36 : good (23.0)
        age > 36 : bad (24.0)
    property = CarOther :
        marital = MaleSingle : good (44.0)
        marital = FemaleDivorsedMarried : bad (28.0)
        marital = MaleDivorsed : good (8.0)
        marital = MaleMarried : bad (12.0)
```

Figure 5. The set of rules for credit scoring

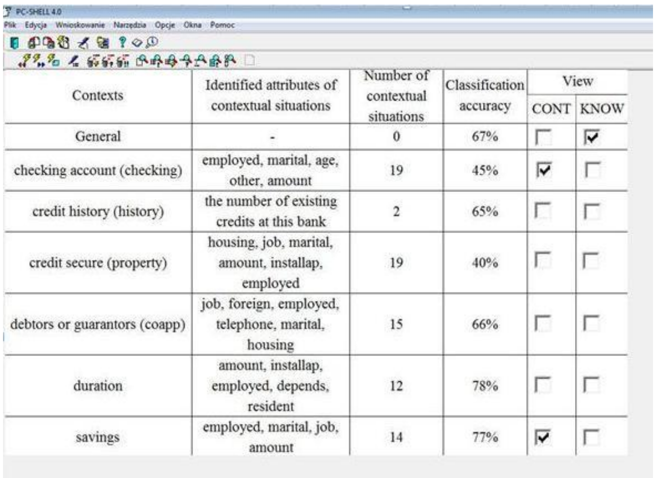
Each descriptor of ‘good’ or ‘bad’ applicants is investigated whether it is context dependent. There is determined the relation between them and the remaining attributes to find up the possible contexts. In this case all descriptors are context dependent, so the number of discovered context is six. Then for each identified context there is conducted selection of appropriate data form data base and is generated context dependent knowledge model. This step is automatic but the user can view either the models or the contexts.

5.2 The problem solving

The system starts with the window for entering the description of problem. If the contexts exist the list of required attribute value is longer. In analyzed case we have six context dependent attributes and thirteen contextual attributes representing for example: customer property and savings, his marital status, housing, employing forms, job characteristics and the like. The list of required attributes values may vary from session to session as knowledge base is changing according to new cases accumulated in data base. The attributes describe either problem solving or its possible contexts.

After entering the required attribute values, a list of discovered contexts for credit scoring is presented (see Fig. 6). The names of contexts are the same as names of context dependent attributes. Each context has a list of its descriptors in the significance order. The assessment of context consists of complexity measure (e.g) number of contextual situations and quality measure (classification accuracy of context dependent knowledge.

Aside from that there is description of decision tree model for problem under consideration (general). The user at this stage may choose general model as solution if the quality is acceptable for him. This the simplest case – version I. More often there are no single model solution for difficult problems as in this case where the quality of general model is too low – 67%. So the user in the solution process evaluates each context and context depended knowledge according to his knowledge and expectation and objective measures.



Contexts	Identified attributes of contextual situations	Number of contextual situations	Classification accuracy	View	
				CONT	KNOW
General	-	0	67%	☐	☑
checking account (checking)	employed, marital, age, other, amount	19	45%	☑	☐
credit history (history)	the number of existing credits at this bank	2	65%	☐	☐
credit secure (property)	housing, job, marital, amount, installap, employed	19	40%	☐	☐
debtors or guarantors (coapp)	job, foreign, employed, telephone, marital, housing	15	66%	☐	☐
duration	amount, installap, employed, depends, resident	12	78%	☐	☐
savings	employed, marital, job, amount	14	77%	☑	☐

Figure 6. The list of identified contexts

Let’s look for example at two specified context: savings and checking account. The user can compare the context attributes that create different contexts. The first two descriptors for these contexts are the same i.e. period of present employment and marital status. Thus maybe choose one of them?. A user can compare contexts not only on the list basis but also may have a decision trees view of pointed contexts. Before decision making user may look for more detailed view on these groups of applicants (Fig. 7).

The context checking account makes great fragmentation with no clear description of applicants that have given type of account in a bank. So user decide to omit this context from further consideration. To well-thought-out the problem under consideration, the context’s view may be insufficient and user may have a look at context dependent knowledge referring to given context. A user can choose one or multiple contexts.

The single choice finishes problem solution as version I. If the user make multiple choice there is a need to decide about the voting schemata for fusion intermediate solutions.

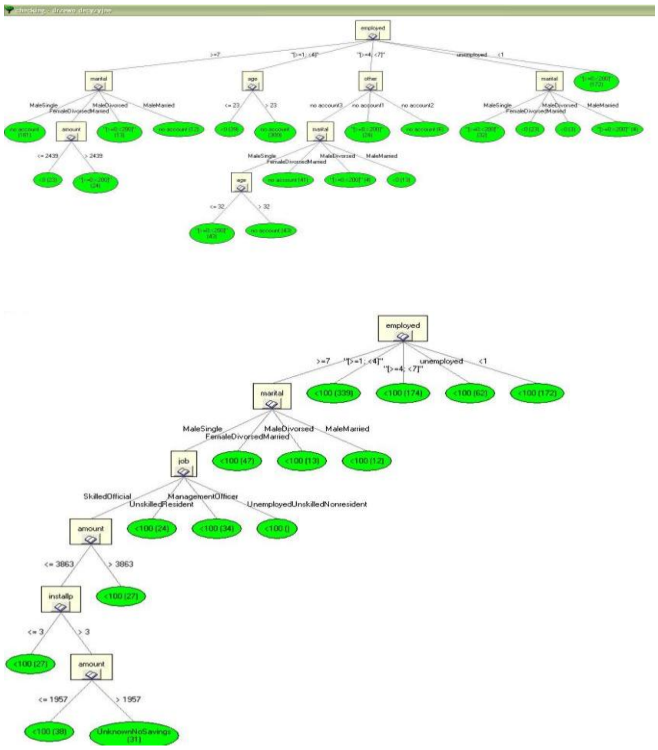
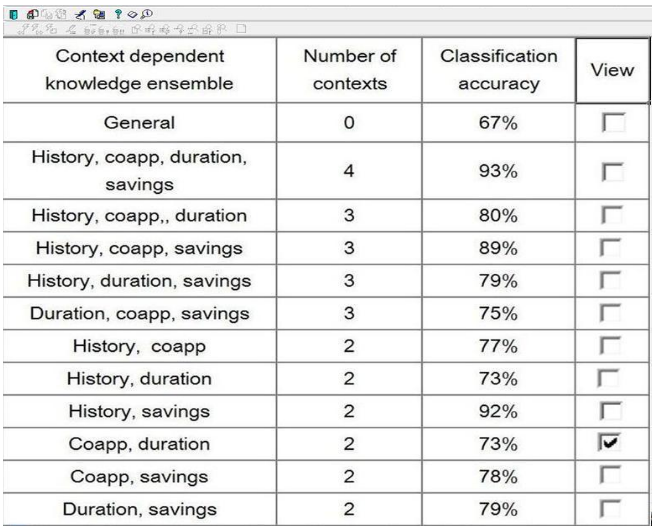


Figure 7. The structure of two contexts: savings and checking

As the results of user analysis the four contexts are chosen to create the solution: credit history, debtors or guarantors (coapp), duration and saving. The possible contextual ensembles are presented on the Fig. 8. There are all possible combination of chosen contexts. The number of possible ensembles for credit scoring is eleven.

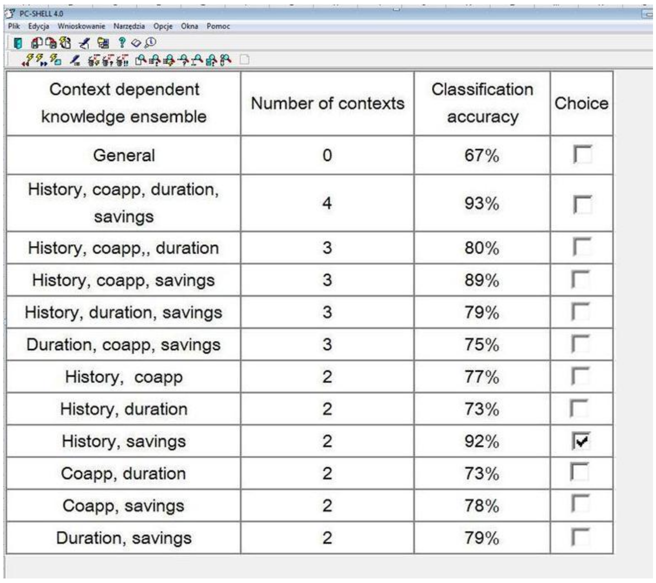


Context dependent knowledge ensemble	Number of contexts	Classification accuracy	View
General	0	67%	☐
History, coapp, duration, savings	4	93%	☐
History, coapp,, duration	3	80%	☐
History, coapp, savings	3	89%	☐
History, duration, savings	3	79%	☐
Duration, coapp, savings	3	75%	☐
History, coapp	2	77%	☐
History, duration	2	73%	☐
History, savings	2	92%	☐
Coapp, duration	2	73%	☑
Coapp, savings	2	78%	☐
Duration, savings	2	79%	☐

Figure 8. The list of possible contextual ensembles

As it can be seen the level of quality for each contextual ensemble is pretty high and is much higher than for general model. Before the final choice the user can view some of propositions listed.

At this stage the user can select acceptable solution (Fig. 9). This is the case of version II.



Context dependent knowledge ensemble	Number of contexts	Classification accuracy	Choice
General	0	67%	☐
History, coapp, duration, savings	4	93%	☐
History, coapp,, duration	3	80%	☐
History, coapp, savings	3	89%	☐
History, duration, savings	3	79%	☐
Duration, coapp, savings	3	75%	☐
History, coapp	2	77%	☐
History, duration	2	73%	☐
History, savings	2	92%	☑
Coapp, duration	2	73%	☐
Coapp, savings	2	78%	☐
Duration, savings	2	79%	☐

Figure 9. The list of the solution choice

The user choice is marked solution (Fig. 9), i.e. ensemble of context dependent knowledge ensemble that includes credit history context and saving context. As we can see there is another solution with higher quality (93%) but more complex and the user skips it. In such cases the choices are not obvious, so the user knowledge and experiences are not overestimated.

The variant III takes place when single context has not acceptable accuracy i.e. below 55%. It means that there is impossible to create any ensemble with the quality higher than random choice. The user has to look at identified contexts more profoundly. Let's look at contexts that were skipped by the user (Table 1).

Table 1. The contexts with low quality

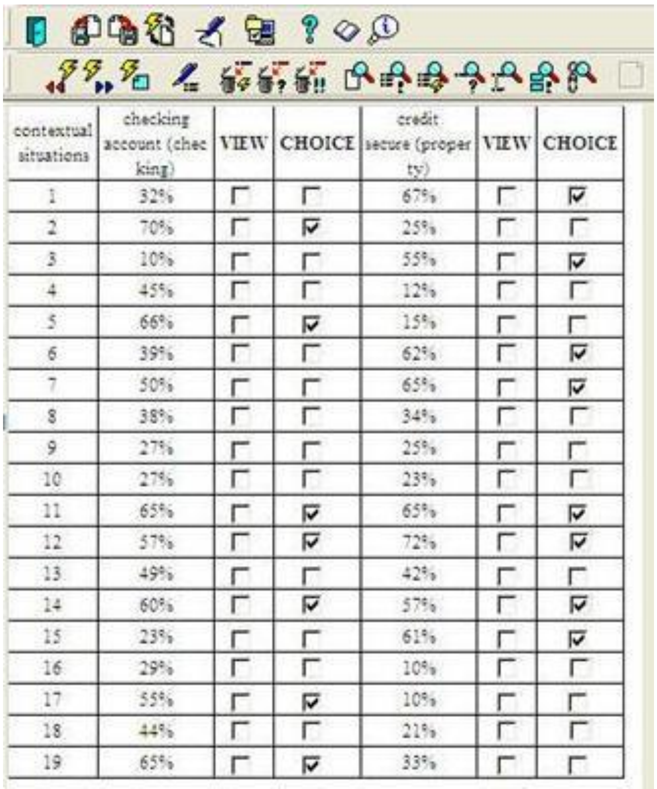
Identified attributes of contextual situations	Number of contextual situations	Classification accuracy
employed, marital, age, other, amount	19	45%
housing, job, marital, amount, installap, employed	19	40%

There are two contexts with 19 contextual situations that required more detailed analysis. The user looks for contextual situations that describe problem under consideration with acceptable quality but on the more detail level (Fig. 10)

As it can be seen in both contexts are contextual situations that can meet user requirements about quality. As in earlier version the user can view each contextual situation and referring context dependent knowledge and evaluate them according to his experience and knowledge.

The user can analyze acceptable contextual situation according to the level of classification accuracy. For example in the context checking account the two groups of applicants have good quality context dependent knowledge i.e. contextual situation no 2 with 70% and contextual situation number 5 with 66%.

The former describe the applicants that are employed for less than one year period and have good account. The latter contextual situation concerns the applicants with none account that are single



contextual situations	checking account (checking)	VIEW	CHOICE	credit secure (property)	VIEW	CHOICE
1	32%	□	□	67%	□	☑
2	70%	□	☑	25%	□	□
3	10%	□	□	55%	□	☑
4	45%	□	□	12%	□	□
5	66%	□	☑	15%	□	□
6	39%	□	□	62%	□	☑
7	50%	□	□	65%	□	☑
8	38%	□	□	34%	□	□
9	27%	□	□	25%	□	□
10	27%	□	□	23%	□	□
11	65%	□	☑	65%	□	☑
12	57%	□	☑	72%	□	☑
13	49%	□	□	42%	□	□
14	60%	□	☑	57%	□	☑
15	23%	□	□	61%	□	☑
16	29%	□	□	10%	□	□
17	55%	□	☑	10%	□	□
18	44%	□	□	21%	□	□
19	65%	□	☑	33%	□	□

Figure 3. The list of contextual situations for two contexts

man and are employed for more than 7 year. The number of applicants described by analyzed contextual situation is also important information for the user The sparse groups sometimes can be interesting but seldom.

The user after profound investigation has chosen seven contextual situations for checking context and eight for property context. Because the chosen context dependent knowledge chunks do not cover all cases included in data base there has to be added general context dependent knowledge. After selecting appropriate voting schemata the contextual ensemble were created. The quality of such solution is very high and equals 83%, that seems exceptionally good solution under these circumstances.

The first results of contextual knowledge handled by en expert system are promising. The decision accuracy was significantly higher than by using general knowledge base, in the range from 4% - 20-%. The users have built different solution and have learnt new relation between contexts, between context dependent knowledge.

6 DISCUSSION ON THE ISSUES RAISED

The keys issues essential in this research include: (1) automatic contextual and context dependent knowledge acquisition, (2) the way of employing the context in problem solving and (3) the role of the user in the process of solution searching.

The *automatic contextual and context dependent knowledge acquisition* requires the contexts to be inside the data base. Thus the contexts are unknown and have to be identified. Such approach gives the possibility to discover new unknown and more complex contexts, to see the problem through multiply contexts and the possibility to more profound look at the problem under consideration It do not means that our proposition is limited to unknown contexts. When the contexts are known the process of context identification can be just omitted.

The most of researchers deal with known and clear defined contexts. In such cases the contexts are just the single concepts to mention just a two. For example, for the word recognition the contexts refer to methods used to solve this problem: sign recognition method, the word recognition by segmentation and the shape of word analysis method[16].

Aspers P.[17] in the contemporary economy refers the contexts to network of actors, arenas of aesthetic creative workers and final consumer markets.

The more complex contexts, besides the more profound insight into the problem under consideration, may cause an disadvantage. This concerns the context using as a way of knowledge partition in knowledge base. The knowledge fragmentation may be too high so the searching for problem solution may be more complicated for the user. This will be the subject of our future research.

The *way of problem solving* by invoking the appropriate contextual knowledge chunks to the current context are common for many researches. We extend it by the possibility to perceive the problem in multiply contexts. The user has the possibility to create many solutions but to do it well he have to be an expert in problem domain. The most related to our work approach, already was mentioned the paper of Ho T.K., Hull J.J., Srihari S.N. They see a solution of the problem of word recognition as an ensemble of three chunks of context dependent knowledge that are generated by the three classifiers each of them in different context. The contextual chunks of contextual knowledge are combined using some schemata of fusion. This differ from our approach arbitrary and automatically set up contexts and contextual knowledge without any user involved in process of problem solving.

The last key issue concerns the *role of an user*. The evolution of intelligent systems is directing to achieve the balance between the user and the system, so that the system should play the role of human's assistant in knowledge acquisition from the user[18]. In our work we have focused on the user role as his active contribution to the problem solving and we omit, for now the utilization of user knowledge to enrich the knowledge base. The reason is the possible frequency of knowledge base modification which can be high, thus such an effort may not be effective. But for some domain it will be possible to work, so it will be our future direction. It is worth mentioning about the many possible solution of the problem which may be composed by the user. This gives the possibility to find the appropriate solution but also to make the user more profound acquainted with the problem and problem limitations.

Basically all the discussed issues related to contextual knowledge refer to decision making processes. We should be

[16] Ho T.K., Hull J.J., Srihari S.N.: Word recognition with multilevel contextual knowledge. Proceedings of the First International Conference on Document Analysis and Recognition 1991 September 30 – October 2, University of Salford

[17] Aspers P.: Contextual Knowledge, *Current Sociology*, September 2006/54; 745-763, http://www.soa4all.eu/contextmanagement.html

[18] Brézillon P.: Context-Based Intelligent Assistant Systems: A discussion based on the Analysis of Two Projects. Proceedings of the 36th Hawaii International Conference on System Sciences – 2003

conscious of the context importance in many other areas. For example contexts are very important in information systems embracing law aspects (different conditions of law enforcement) or polymorphic ways of applying distinct procedures for particular environmental sides. As a result organizational knowledge get multidimensional and contextual dependent characteristics. Integration of different contextual aspects organizational knowledge becomes a big challenge for future research.

REFERENCES

[1] Akman V., Rethinking context as a social construct, „Journal of Pragmatics” 2000, Vol. 32, No. 6, pp. 743-759. Aspers P.: Contextual Knowledge, *Current Sociology*, September 2006/54;745-763, http://www.soa4all.eu/contextmanagement.html

[2] Boy G., Gruber T.R. (1990): Intelligent Assistant Systems: Support for Integrated Human-Machine Systems. AAAI Spring Symposium on *Knowledge-Based Human-Computer Communication*, March 1990, Stanford University. Brézillon P. J.: Conceptualized explanations. International Conference on Expert Systems for Development, Bangkok, Thailand, March 1994

[3] Brézillon P., Pomerol J.-Ch., Contextual knowledge sharing and cooperation in intelligent assistant systems, „Le Travail Humain” 1999, Vol. 62, No. 3, pp. 223-246.

[4] Brezillon P., Some characteristics of context, Lecture Notes in Computer Science, 2006, Vol. 4031, pp. 146-154.

[5] Brézillon, P. (2011) "From expert systems to context-based intelligent assistant systems: a testimony". The Knowledge Engineering Review, 26(1), pp. 19-24.

[6] Cram J., Sayers R. (2001): Creating and Managing Context: The Use of Knowledge Management Principles to Deliver Virtual Information Services to Schools. ASLA XVII Conference, Queensland 2001

[7] Harries, M. B., Sammut, C., Horn, K.: Extracting hidden context, Machine Learning, 32, 1998 Hirst G., Context as a spurious concept, [w:] Paper Presented at the AAAI-97 Fall Symposium on Context in Knowledge Representation and Natural Language, MIT, Cambridge, Mass., 1997.

[8] Ho T.K., Hull J.J., Srihari S.N.: Word recognition with multilevel contextual knowledge. Proceedings of the First International Conference on Document Analysis and Recognition 1991 September 30 – October 2, University of Salford

[9] Jakubczyc J., Contextual classifier ensemble for predicting customer churn, [in:] M. Nycz, M. Owoc (eds.), Knowledge Acquisition and Management, Research Papers of Wrocław University of Economics No. 133, Publishing House of the Wrocław University of Economics, Wrocław 2010, pp.42-57.

[10] Jakubczyc J., Contextual classifier ensembles, [in :] W. Abramowicz (ed.), LNCS 4439 – Business Information Systems, , Springer, Berlin, Heidelberg 2007, pp. 562-569.

[11] Jakubczyc J.A., The context in concept learning, [in:] Owoc M.L., Nycz M. (eds.) Knowledge Acquisition and Management, Research Papers of Wrocław University of Economics No. 25, Publishing House of the Wrocław University of Economics, Wrocław 2008.

Artificial Intelligence for Knowledge Management with BPMN and Rules

Antoni Ligęza, Krzysztof Kluza, Grzegorz J. Nalepa, Weronika T. Adrian[1] and **Tomasz Potempa**[2]

Abstract. This paper presents a framework combining BPMN and BR as a tool for Knowledge Management. An attempt at providing a common model supported with AI techniques and tools is put forward. Through an extended example it is shown how to combine BPMN and BR and how to pass to semantic level enabling building executable specifications and knowledge analysis. Some of the problems concerning these two approaches can be to certain degree overcome thanks to their complementary nature. We only deal with a restricted view of Knowledge Management, where knowledge can be modeled explicitly in a formal representation, and it does not take into account the knowledge residing in people's heads.

1 INTRODUCTION

Design, development and analysis of progressively more and more complex business processes require advanced methods and tools. Apart from variety of classical AI stuff, two generic modern approaches to modeling such processes have recently gained wider popularity: *Business Process Model and Notation* (BPMN) [23] and *Business Rules* (BR) [1, 4]. Although aimed at a common target, both of these approaches are rather mutually complementary and offer distinctive features enabling process modelling.

BPMN constitutes a set of graphical symbols, such as links modeling workflow, various splits and joins, events and boxes representing data processing activities. It is a transparent visual tool for modeling complex processes promoted by OMG [23]. What is worth underlying is the expressive power of current BPMN. In fact it allows for modeling conditional operations, loops, event-triggered actions, splits and joins of data flow paths and communication processes. Moreover, modeling can take into account several levels of abstraction enabling a hierarchical approach.

BPMN can be considered as *procedural knowledge representation*; a BPMN diagram represents in fact a set of interconnected procedures. Although BPMN provides transparent, visual representation of the process, due to lack of formal model semantics it makes attempts at more rigorous analysis problematic. Further, even relatively simple inference requires a lot of space for representation; there is no easy way to specify declarative knowledge, e.g. in the form of rules.

Business Rules, also promoted by OMG [21, 22], offer an approach to specification of knowledge in a *declarative* manner. The way the rules are applied is left over to the user when it comes to rule execution. Hence, rules can be considered as *declarative knowledge specification*; inference control is not covered by basic rules.

These two approaches are to certain degree complementary: BR provide declarative specification of domain knowledge, which can be encoded into a BPMN model. On the other hand, BPMN diagram can be used as procedural specification of the workflow, including inference control [7]. However, BPMN lacks of a *formal declarative model* defining the semantics and logic behind the diagram. Hence, defining and analyzing correctness of BPMN diagrams is a hard task. There are papers undertaking the issues of analysis and verification of BPMN diagrams [3, 9, 24, 26]. However, the analysis is performed mostly at the *structural* level and does not take into account the semantics of dataflow and control knowledge.

In this position paper, we follow the ideas initially presented in [11]. An attempt at defining foundations for a more formal, logical, declarative model of the most crucial elements of BPMN diagrams combined with BR is undertaken. We pass from logical analysis of BPMN component to their logical models, properties and representation in PROLOG [12]. The model is aimed at enabling definition and further analysis of selected formal properties of a class of restricted BPMN diagrams. The analysis should take into account properties constituting reasonable criteria of correctness. The focus is on development of a formal, declarative model of BPMN components and its overall structure. In fact, a combination of the recent approaches to development and verification of rule-based systems [13, 17, 19] seems to have potential influence on the BPMN analysis.

2 MOTIVATION

Knowledge has become a valuable resource and a decisive factor for successful operation of organizations, companies and societies. As vast amounts of knowledge are in use, tools supporting Knowledge Management (KM) are inevitable support for Decision Makers. Such tools can be classified into the following categories:

- *Conceptual Models* — various symbolic and visual ways of Knowledge Representation (KR), analysis, and supporting design of knowledge-intensive systems and applications; as an example one can mention various schemes, graphs, networks and diagrams, with UML [18] and BPMN [6] being some perfect examples,
- *Logical Models* — more formal KR and knowledge processing (reasoning, inference) tools, supporting both *representation* and *application* of knowledge [2, 15]. It is important that such models typically support also *semantic* issues; as an example one can mention various types of logics and logic-derived formalisms including rules and Business Rules (BR) as some perfect examples.
- *Functional and Procedural Models* — these include all algorithmic-type recipes for performing operations; some typical examples may vary from linguistically represented procedures, e.g. ISO, to programs encoded with any programming languages.

[1] AGH University of Science and Technology, Krakow, Poland, email: {ligeza,kluza,gjn,wta}@agh.edu.pl

[2] Higher School of Tarnów, Tarnow, Poland, email: tpotempa@gmail.com

When speaking about Conceptual Models one usually assumes more or less informal, abstract, illustrative presentation of concepts, relations, activities, etc. In case of Logical Models, clear syntax and semantic rules are in background; this assures possibility of identification and verification of properties, such as (i) *consistency*, (ii) *completeness*, (iii) *unique interpretation* (lack of ambiguity), (iv) *efficiency* (minimal representation, lack of redundancy, efficient operation), (v) *processability*. Some further requirements may refer to: readability and transparency, easy modifications and extensionability, support for knowledge acquisition and encoding, etc.

The above-mentioned models are used to represent, analyze, process and optimize knowledge. Note that there are at least the following types of knowledge aimed at separate goals and requiring different way of processing:

- *typological* or *taxonometric* knowledge (e.g. a taxonomy in typed logics and languages or TBox in Description Logics),
- *factographic* knowledge representing facts and relations about object (e.g. a set of the FOPC atomic formulae or ABox in DL),
- *inferential* or *transformation* knowledge — specification of legal knowledge rewriting rules or production rules,
- *integrity and constraints* knowledge on what is impossible, not allowed, etc.
- *meta-knowledge* — all about how to use the basic knowledge (e.g. inference control rules).

Now, most typical KM activities require solving such issues as:

1. Knowledge Representation,
2. Inference — knowledge processing rules,
3. Inference Control — principles on how to apply inference rules in a correct and efficient manner,
4. Knowledge Acquisition and Updating,
5. Knowledge Analysis and Verification,
6. Friendly User Interface,
7. Generalization and Learning.

A tool, or a set of tools, for efficient Knowledge Management should support as many KM activities in a smooth way and deserve handling as many types of knowledge within a single framework.

2.1 BPMN as a tool for KM

BPMN [23] appears to be an effective choice for Knowledge Management tasks. It offers a wide spectrum of graphical elements for visualizing of events, activities, workflow splits and merges, etc. It can be classified as Conceptual Modeling tool of high expressive power, practically useful and still readable to public.

Let us briefly analyze the strengths and weaknesses of BPMN as a KM tool. It is mostly a way of *procedural knowledge specification*, so it supports p. 3 above, but neither p. 1 nor 2. Certainly, refering to p. 6 its user interface is nice. An important issue about BPMN is that it covers three important aspects of knowledge processing:

- *inference control* or *workflow control*, including diagrammatic specification of the process with partial ordering, switching and merging of flow,
- *data processing* or *data flow* specification, including input, output and internal data processing,
- *structural representation* of the process as a whole, allowing for visual representation at several levels of hierarchy.

Some more serious weakness issues concerning characteristics and activities presented in Section 2 are as follows:

- BPMN — being a Conceptual Modeling tool — does not provide formal semantics,
- it is inadequate for knowledge analysis and verification (p. 5),
- it neither support declarative representation of taxonometric, factographical, nor integrity knowledge.

However, some of these weaknesses can be overcome by combination of BPMN with Business Rules.

2.2 Business Rules as a tool for KM

Business Rules (BR) can be classified as Logical Model for KR. They constitute a declarative specification of knowledge. There can be different types of BR serving different purposes; in fact all the types of knowledge (taxonometric, factographical, transformation, integrity, and meta) can be encoded with BR.

A closer look at foundations of Rule-Based Systems [10] shows that rules can:

- have high expressive power depending on the logic in use,
- provide elements of procedural control,
- undergo formal analysis.

Rules, especially when grouped into decision modules (such as decision tables) [14], are easier to analyze. However, the possibility of analysis depends on the accepted *knowledge representation language*, and in fact – the logic in use. Formal models of rule-based systems and analysis issues are discussed in detail in [10].

The main weakness of BR consists in lack of procedural (inference control) specification and transparent knowledge visualization. However, these issues can be solved at the BPMN level.

2.3 BPMN and BR: Toward an Integration Framework

In order to integrate BPMN and BR, a framework combining and representing intrinsic mechanisms of these two approaches is under development. It should be composed of the following elements:

- Workflow Structure/Sequence Graph (WSG) — an AND-OR graph representing a workflow structure at abstract level,
- Logical Specification of Control (LSC) — logical labels for WSG,
- Dataflow Sequence Graph (DSG) — a DFD-type graph showing the flow of data,
- Logical Specification of Data (LSD) — constraints imposed on data being input, output or processed at some nodes.

3 WORKFLOW STRUCTURAL GRAPH for BPMN

Workflow Structure/Sequence Graph (WSG) is in fact a simplified structural model of BPMN diagrams. It constitutes a restricted abstraction of crucial intrinsic workflow components. As for events, only start and termination events are taken into account. Main knowledge processing units are activities (or tasks). Workflow control is modeled by two subtypes of gateways: split and join operations. Finally, workflow sequence is modeled by directed links. No time or temporal aspect is considered.

The following elements will be taken into consideration [11]:

- $\mathbb{S}$ — a non-empty set of *start events* (possibly composed of a single element),
- $\mathbb{E}$ — a non-empty set of *end events* (possibly composed of a single element),
- $\mathbb{T}$ — a set of *activities* (or *tasks*); a task $T \in \mathbb{T}$ is a finite process with single input and single output, to be executed within a finite interval of time,
- $\mathbb{G}$ — a set of *split gateways* or *splits*, where branching of the workflow takes place; three disjoint subtypes of splits are considered:

- $\mathbb{GX}$ — a set of *exclusive splits* where one and only one alternative path can be followed (a split of the $EX - OR$ type),
- $\mathbb{GP}$ — a set of *parallel splits* where all the paths of the workflow are to be followed (a split of the AND type or a *fork*), and
- $\mathbb{GO}$ — a set of *inclusive splits* where one or more paths should be followed (a split of the OR type).

- $\mathbb{M}$ — a set of *merge gateways* or *joins*, where two or more paths meet; three disjoint subtypes of merge (join) nodes are considered:
  - $\mathbb{MX}$ — a set of *exclusive merge* nodes where one and only one input path is taken into account (a merge of the $EX - OR$ type),
  - $\mathbb{MP}$ — a set of *parallel merge* nodes where all the paths are combined together (a merge of the AND type), and
  - $\mathbb{MO}$ — a set of *inclusive merge* nodes where one or more paths influence the subsequent item (a merge of the OR type).
- $\mathbb{F}$ — a set of workflow links, $\mathbb{F} \subseteq \mathbb{O} \times \mathbb{O}$, where $\mathbb{O} = \mathbb{S} \cup \mathbb{E} \cup \mathbb{T} \cup \mathbb{G} \cup \mathbb{M}$ is the join set of objects. All the component sets are pairwise disjoint.

The splits and joins depend on logical conditions assigned to particular branches. It is assumed that there is defined a partial function $\mathtt{Cond}\colon \mathbb{F} \to \mathbb{C}$ assigning logical formulae to links. In particular, the function is defined for links belonging to $\mathbb{G} \times \mathbb{O} \cup \mathbb{O} \times \mathbb{M}$, i.e. outgoing links of split nodes and incoming links of merge nodes. The conditions are responsible for workflow control. For intuition, an exemplary simple BPMN diagram is presented in Fig. 1.

In order to assure *structural correctness* of BPMN diagrams a set of restrictions on the overall diagram structure is typically defined; they determine the so-called *well-formed diagram* [24]. Classical AI graph search methods can be applied for analysis. However, a well-formed diagram does not assure that for any input knowledge the process can be executed leading to a (unique) solution. This depends on the particular input data, its transformation during processing, correct work of particular objects, and correct control defined by the branching/merging conditions assigned to links.

The further issues, i.e. Logical Specification of Control (LSC), Dataflow Sequence Graph (DSG), and Logical Specification of Data (LSD) will be analyzed on the base of an example presented below.

4 BPMN AND BR EXAMPLE: THE THERMOSTAT CASE

In order to provide intuitions, the theoretical considerations will be illustrated with a simple exemplary process. The process goal is to establish the so-called *set-point* temperature for a thermostat system [20]. The selection of the particular value depends on the season, whether it is a working day or not, and the time of the day.

Consider the following set of declarative rules specifying the process. There are eighteen inference rules (production rules):

Rule 1: $aDD \in \{\mathit{monday}, \mathit{tuesday}, \mathit{wednesday}, \mathit{thursday}, \mathit{friday}\} \longrightarrow aTD = wd.$

Rule 2: $aDD \in \{\mathit{saturday}, \mathit{sunday}\} \longrightarrow aTD = wk.$

Rule 3: $aTD = wd \wedge aTM \in (9, 17) \longrightarrow aOP = dbh.$

Rule 4: $aTD = wd \wedge aTM \in (0, 8) \longrightarrow aOP = ndbh.$

Rule 5: $aTD = wd \wedge aTM \in (18, 24) \longrightarrow aOP = ndbh.$

Rule 6: $aTD = wk \longrightarrow aOP = ndbh.$

Rule 7: $aMO \in \{\mathit{january}, \mathit{february}, \mathit{december}\} \longrightarrow aSE = sum.$

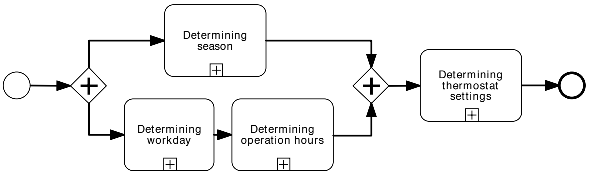


Figure 1. An example BPMN diagram — top-level specification of the thermostat system

Rule 8: $aMO \in \{\mathit{march}, \mathit{april}, \mathit{may}\} \longrightarrow aSE = aut.$

Rule 9: $aMO \in \{\mathit{june}, \mathit{july}, \mathit{august}\} \longrightarrow aSE = win.$

Rule 10: $aMO \in \{\mathit{september}, \mathit{october}, \mathit{november}\} \longrightarrow aSE = spr.$

Rule 11: $aSE = spr \wedge aOP = dbh \longrightarrow aTHS = 20.$

Rule 12: $aSE = spr \wedge aOP = ndbh \longrightarrow aTHS = 15.$

Rule 13: $aSE = sum \wedge aOP = dbh \longrightarrow aTHS = 24.$

Rule 14: $aSE = sum \wedge aOP = ndbh \longrightarrow aTHS = 17.$

Rule 15: $aSE = aut \wedge aOP = dbh \longrightarrow aTHS = 20.$

Rule 16: $aSE = aut \wedge aOP = ndbh \longrightarrow aTHS = 16.$

Rule 17: $aSE = win \wedge aOP = dbh \longrightarrow aTHS = 18.$

Rule 18: $aSE = win \wedge aOP = ndbh \longrightarrow aTHS = 14.$

Let us briefly explain these rules. The first two rules define if we have today (aTD) a workday (wd) or a weekend day (wk). Rules 3-6 define if the operation hours (aOP) are during business hours (dbh) or not during business hours ($ndbh$); they take into account the workday/weekend condition and the current time (hour). Rules 7-10 define the season (aSE) is summer (sum), autumn (aut), winter (win) or spring (spr). Finally, rules 11-18 define the precise setting of the thermostat ($aTHS$). Observe that the set of rules is flat; basically no control knowledge is provided.

Now, let us attempt to visualize a business process defined with these rules. A BPMN diagram of the process is presented in Fig. 1. After start, the process is split into two independent paths of activities. The upper path is aimed at determining the current season (aSE; it can take one of the values $\{sum, aut, win, spr\}$; the detailed specification is provided with rules 7-10). A visual specification of this activity with an appropriate set of rules is shown in Fig. 2.

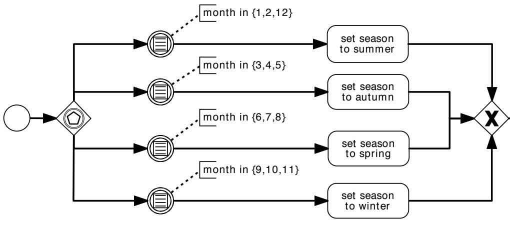


Figure 2. An example BPMN diagram — detailed specification a BPMN task

The lower path determines whether the day (aDD) is a workday ($aTD = wd$) or a weekend day ($aTD = wk$), both specifying the value of today (aTD; specification provided with rules 1 and 2), and then, taking into account the current time (aTM), whether the operation (aOP) is during business hours ($aOP = dbh$) or not ($aOP = ndbh$); the specification is provided with rules 3-6. This is illustrated with Fig. 3 and Fig. 4.

Finally, the results are merged together, and the final activity consists in determining the thermostat settings ($aTHS$) for particular

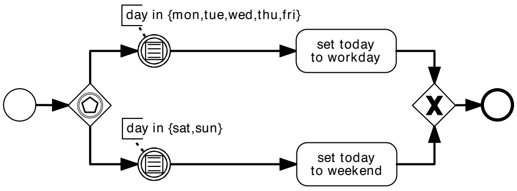


Figure 3. An example BPMN diagram — detailed specification of determining the day task

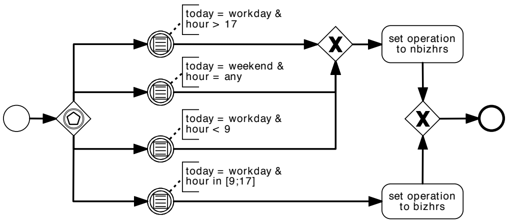

today = workday & hour > 17
today = weekend & hour = any
today = workday & hour < 9
today = workday & hour in [9;17]
set operation to nbizhrs
set operation to bizhrs

Figure 4. An example BPMN diagram — detailed specification of working hours task

season (aSE) and time (aTM) (the specification is provided with rules 11-18). This is illustrated with Fig. 5.

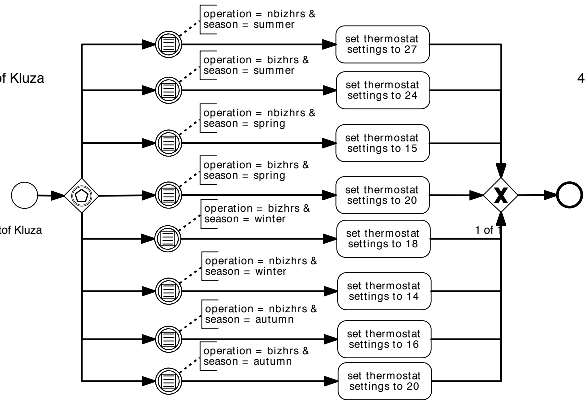

operation = nbizhrs & season = summer
operation = bizhrs & season = summer
operation = nbizhrs & season = spring
operation = bizhrs & season = spring
operation = bizhrs & season = winter
operation = nbizhrs & season = winter
operation = nbizhrs & season = autumn
operation = bizhrs & season = autumn
set thermostat settings to 27
set thermostat settings to 24
set thermostat settings to 15
set thermostat settings to 20
set thermostat settings to 18
set thermostat settings to 14
set thermostat settings to 16
set thermostat settings to 20

Figure 5. An example BPMN diagram — detailed specification of the final thermostat setting task

Even in this simple example, answers to the following important questions are not obvious:

1. *data flow correctness*: Is any of the four tasks/activities specified in a correct way? Will each task end with producing desired output for any admissible input data?
2. *split consistency*: Will the workflow possibly explore all the paths after a split? Will it always explore at least one?
3. *merge consistency*: Will it be always possible to merge knowledge coming from different sources at the merge node?
4. *termination/completeness*: Does the specification assure that the system will always terminate producing some temperature specification for *any* admissible input data?
5. *determinism*: Will the output setting be determined in a unique way?

Note that we do not ask about *correctness* of the result; in fact, the rules embedded into a BPMN diagram provide a kind of *executable specification*, so there is no reference point to claim that final output is correct or not.

5 LOGICAL SPECIFICATION OF CONTROL

This section is devoted to analysis of logical specification of control. In fact, two types of control elements are analyzed: split nodes and merge nodes.

5.1 Analysis of Split Conditions

An exclusive split $GX(q_1, q_2, \ldots q_k) \in \mathbb{GX}$ with k outgoing links is modelled by a fork structure assigned excluding alternative of the form:

$$q_1 \veebar q_2 \veebar \ldots \veebar q_k,$$

where $q_i \wedge q_j$ is always false for $i \neq j$. An exclusive split can be considered correct if and only if at least one of the alternative conditions is satisfied. We have the following logical requirement:

$$\models q_1 \vee q_2 \vee \ldots \vee q_k, \tag{1}$$

i.e. the disjunction is in fact a tautology. In practice, to assure (1), a predefined exclusive set of conditions is completed with a default q_0 condition defined as $q_0 = \neg q_1 \wedge \neg q_2 \wedge \ldots \wedge \neg q_k$; obviously, the formula $q_0 \vee q_1 \vee q_2 \vee \ldots \vee q_k$ is a tautology.

Note that in case when an input restriction formula ϕ is specified, the above requirement given by (1) can be relaxed to:

$$\phi \models q_1 \vee q_2 \vee \ldots \vee q_k. \tag{2}$$

An inclusive split $GO(q_1, q_2, \ldots q_k) \in \mathbb{GO}$ is modelled as disjunction of the form:

$$q_1 \vee q_2 \vee \ldots \vee q_k,$$

An inclusive split to be considered correct must also satisfy formula (1), or at least (2). As before, this can be achieved through completing it with the q_0 default formula.

A parallel split $GP(q_1, q_2, \ldots q_k) \in \mathbb{GP}$ is referring to a fork-like structure, where all the outgoing links should be followed in any case. For simplicity, it can be considered as an inclusive one, where all the conditions assigned to outgoing links are set to *true*.

Note that, if ϕ is the restriction formula valid for data at the input of the split, then any of the output restriction formula is defined as $\phi \wedge q_i$ for any of the outgoing link i, $i = 1, 2, \ldots, k$.

5.2 Analysis of Merge Conditions

Consider a workflow merge node, where k knowledge inputs satisfying restrictions $\phi_1, \phi_2, \ldots, \phi_k$ respectively meet together, while the selection of particular input is conditioned by formulae $p_1, p_2, \ldots, p_k$, respectively.

An exclusive merge $MX(p_1, p_2, \ldots, p_k) \in \mathbb{MX}$ of k inputs is considered correct if and only if the conditions are pairwise disjoint, i.e.

$$\not\models p_i \wedge p_j \tag{3}$$

for any $i \neq j$, $i, j \in \{1, 2, \ldots, k\}$. Moreover, to assure that the merge works, at least one of the conditions should hold:

$$\models p_1 \vee p_2 \vee \ldots \vee p_k, \tag{4}$$

i.e. the disjunction is in fact a tautology. If the input restrictions $\phi_1, \phi_2, \ldots, \phi_k$ are known, condition (4) might possibly be replaced by $\models (p_1 \wedge \phi_1) \vee (p_2 \wedge \phi_2) \vee \ldots \vee (p_k \wedge \phi_k)$.

Note that in case a join input restriction formula ϕ is specified, the above requirement can be relaxed to:

$$\phi \models p_1 \vee p_2 \vee \ldots \vee p_k, \tag{5}$$

and if the input restrictions $\phi_1, \phi_2, \ldots, \phi_k$ are known, it should be replaced by $\phi \models (p_1 \wedge \phi_1) \vee (p_2 \wedge \phi_2) \vee \ldots \vee (p_k \wedge \phi_k)$.

An inclusive merge $MO(p_1, p_2, \ldots, p_k) \in \mathbb{MO}$ of k inputs is considered correct if one is assured that the merge works — condition (4) or (5) hold.

A parallel merge $MP \in \mathbb{MP}$ of k inputs is considered correct by default. However, if the input restrictions $\phi_1, \phi_2, \ldots, \phi_k$ are known, a consistency requirement for the combined out takes the form that ϕ must be consistent (satisfiable), where:

$$\phi = \phi_1 \wedge \phi_2 \wedge \ldots \wedge \phi_k \tag{6}$$

An analogous requirement can be put forward for the active links of an inclusive merge.

$$\models p_1 \wedge p_2 \wedge \ldots \wedge p_k, \tag{7}$$

i.e. the conjunction is in fact a tautology, or at least

$$\phi \models p_1 \wedge p_2 \wedge \ldots \wedge p_k. \tag{8}$$

In general, parallel merge can be made correct in a trivial way by putting $p_1 = p_2 = \ldots = p_k = true$.

Note that even correct merge leading to a satisfiable formula assure only passing the merge node; the funnel principle must further be satisfied with respect to the following-in-line object. To illustrate that consider the input of the component determining thermostat setting (see Fig. 1). This is the case of parallel merge of two inputs. The joint formula defining the restrictions on combined output of the components for determining season and determining operation hours is of the form:

$$\phi = (aSE = sum \vee aSE = aut \vee aSE = win \vee \\ aSE = spr) \wedge (aOP = dbh \vee aOP = ndbh).$$

A simple check of all possible combinations of season and operation hours shows that all the eight possibilities are covered by preconditions of rules 11-18; hence the funnel condition (11) holds.

6 DATAFLOW SEQUENCE GRAPH

A Dataflow Sequence Graph (DSG) can be any DFD-type graph showing the flow of data that specifies the data transfers among data processing components. It shows that data produced by certain components should be sent to some next-in-chain ones. In the case of our thermostat example, it happens that the DSG can be represented with the graph shown in Fig. 1 — the workflow and the dataflow structure are the same.

7 LOGICAL SPECIFICATION OF DATA

Logical Specification of Data (LSD) are constraints on data being input, output or processed at some nodes.

7.1 Logical Constraints on Component Behavior

In this section we put forward some minimal requirements defining correct work of rule-based process components performing BPMN activities. Each such component is composed of a set of inference rules, designed to work within the same context; in fact, preconditions of the rules incorporate the same attributes. In our example, we have four such components: determining workday (rules 1-2), determining operation hours (rules 3-6), determining season (rules 7-10) and determining the thermostat setting (rules 11-18).

In general, the outermost logical model of a component T performing some activity/task can be defined as a triple of the form:

$$T = (\psi_T, \varphi_T, \mathcal{A}), \tag{9}$$

where ψ_T is a formula defining the restrictions on the component input, φ_T defines the restrictions for component output, and $\mathcal{A}$ is an algorithm which for a given input satisfying ψ_T produces an (desirably uniquely defined) output, satisfying φ_T. For intuition, ψ_T and φ_T define a kind of a 'logical tube' — for every input data satisfying ψ_T (located at the entry of the tube), the component will produce and output satisfying φ_T (still located within the tube at its output). The precise recipe for data processing is given by algorithm $\mathcal{A}$.

The specification of a rule-based process component given by (9) is considered *correct*, if and only if for any input data satisfying ψ_T the algorithm $\mathcal{A}$ produces an output satisfying φ_T. It is further *deterministic* (unambiguous) if the generated output is unique for any admissible input.

For example, consider the component determining operation hours. Its input restriction formula ψ_T is the disjunction of precondition formulae $\psi_3 \vee \psi_4 \vee \psi_5 \vee \psi_6$, where ψ_i is a precondition formula for rule i. We have $\psi_T = ((aTD = wd) \wedge (aTM \in [0,8] \vee aTM \in [9,17] \vee aTM \in [18,24])) \vee (aTD = wk)$. The output restriction formula is given by $\varphi_T = (aOP = dbh) \vee (aOP = ndbh)$. The algorithm is specified directly by the rules; rules are in fact a kind of *executable specification*.

In order to be sure that the produced output is unique, the following *mutual exclusion* condition should hold:

$$\not\models \psi_i \wedge \psi_j \tag{10}$$

for any $i \neq j$, $i, j \in \{1, 2, \ldots, k\}$. A simple analysis shows that the four rules have mutually exclusive preconditions, and the joint precondition formula ψ_T covers any admissible combination of input parameters; in fact, the subset of rules is locally *complete* and *deterministic* [10].

7.2 Logical Specification of Data Flow

In our example we consider only rule-based components. Let ϕ define the context of operation, i.e. a formula defining some restrictions over the current state of the knowledge-base that must be satisfied before the rules of a component are explored. For example, ϕ may be given by $\varphi_{T'}$ of a component T' directly preceding the current one. Further, let there be k rules in the current component, and let ψ_i denote the joint precondition formula (a conjunction of atoms) of rule i, $i = 1, 2, \ldots, k$. In order to be sure that at least one of the rules will be fired, the following condition must hold:

$$\phi \models \psi_T, \tag{11}$$

where $\psi_T = \psi_1 \vee \psi_2 \vee \ldots \vee \psi_k$ is the disjunction of all precondition formulae of the component rules. The above restriction will be called the *funnel principle*. For intuition, if the current knowledge specification satisfies restriction defined by ϕ, then at least one of the formula preconditions must be satisfied as well.

For example, consider the connection between the component determining workday and the following it component determining operation hours. After leaving the former one, we have that $aTD = wd \vee aTD = wk$. Assuming that the time can always be read as an input value, we have $\phi = (aTD = wd \vee aTD = wk) \wedge aTM \in [0, 24]$. On the other hand, the disjunction of precondition formulae $\psi_3 \vee \psi_4 \vee \psi_5 \vee \psi_6$ is given by $\psi_T = (aTD = wd) \wedge (aTM \in [0, 8] \vee aTM \in [9, 17] \vee aTM \in [18, 24])) \vee aTD = wk$. Obviously, the funnel condition given by (11) holds.

8 CONCLUSIONS AND FUTURE WORK

In this paper, BPMN and BR were explored as tools for Knowledge Management. It is argued that integration of these approaches can overcome some disadvantages of these approaches when considered in separate. Four areas of knowledge specification were put forward: Workflow Specification Graph, Logical Specification of Control (missing in current BPMN), Dataflow Sequence Graph and Logical specification of Data. The ideas of knowledge representation and analysis were illustrated with an example.

The original contribution of our work consists in presenting a framework combining BPMN and BR as a tool for Knowledge Management. The papers only deals with a restricted view of Knowledge Management, where knowledge can be modeled explicitly in a formal representation, and it does not take into account the knowledge residing in people's heads.

As future work, a more complex modeling and verification approach is considered. In the case of modeling issue, we plan to implement this approach by extending one of the existing BPMN tools in order to integrate it with the HeKatE Qt Editor (HQEd) for XTT2-based Business Rules [5]. XTT2 [19] constitutes a formalized attributive language for representing rules in decision tables. Thus, the XTT2 rules (and tables) can be formally analyzed using the so-called verification HalVA framework [6] Although table-level verification can be performed with HalVA [16], the global verification is a more complex issue [8]. Our preliminary works on global verification have been presented in [25].

REFERENCES

[1] S. W. Ambler. Business Rules. http://www.agilemodeling.com/artifacts/businessRule.htm, 2003.

[2] Szymon Bobek, Krzysztof Kaczor, and Grzegorz J. Nalepa, 'Overview of rule inference algorithms for structured rule bases', *Gdansk University of Technology Faculty of ETI Annals*, **18**(8), 57–62, (2010).

[3] Remco M. Dijkman, Marlon Dumas, and Chun Ouyang, 'Formal semantics and automated analysis of BPMN process models. preprint 7115', Technical report, Queensland University of Technology, Brisbane, Australia, (2007).

[4] *Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches*, eds., Adrian Giurca, Dragan Gašević, and Kuldar Taveter, Information Science Reference, Hershey, New York, May 2009.

[5] Krzysztof Kluza, Krzysztof Kaczor, and Grzegorz J. Nalepa, 'Enriching business processes with rules using the Oryx BPMN editor', in *Artificial Intelligence and Soft Computing: 11th International Conference, ICAISC 2012: Zakopane, Poland, April 29–May 3, 2012*, eds., Leszek Rutkowski and [et al.], volume 7268 of *Lecture Notes in Artificial Intelligence*, pp. 573–581. Springer, (2012).

[6] Krzysztof Kluza, Tomasz Maślanka, Grzegorz J. Nalepa, and Antoni Ligęza, 'Proposal of representing BPMN diagrams with XTT2-based business rules', in *Intelligent Distributed Computing V. Proceedings of the 5th International Symposium on Intelligent Distributed Computing – IDC 2011, Delft, the Netherlands – October 2011*, eds., Frances M.T. Brazier, Kees Nieuwenhuis, Gregor Pavlin, Martijn Warnier, and Costin Badica, volume 382 of *Studies in Computational Intelligence*, 243–248, Springer-Verlag, (2011).

[7] Krzysztof Kluza, Grzegorz J. Nalepa, and Łukasz Łysik, 'Visual inference specification methods for modularized rulebases. Overview and integration proposal', in *Proceedings of the 6th Workshop on Knowledge Engineering and Software Engineering (KESE6) at the 33rd German Conference on Artificial Intelligence September 21, 2010, Karlsruhe, Germany*, eds., Grzegorz J. Nalepa and Joachim Baumeister, pp. 6–17, Karlsruhe, Germany, (2010).

[8] Krzysztof Kluza, Grzegorz J. Nalepa, Marcin Szpyrka, and Antoni Ligęza, 'Proposal of a hierarchical approach to formal verification of BPMN models using Alvis and XTT2 methods', in *7th Workshop on Knowledge Engineering and Software Engineering (KESE2011) at the Conference of the Spanish Association for Artificial Intelligence (CAEPIA 2011): November 10, 2011, La Laguna (Tenerife), Spain*, eds., Joaquin Canadas, Grzegorz J. Nalepa, and Joachim Baumeister, pp. 15–23, (2011).

[9] Vitus S. W. Lam, 'Formal analysis of BPMN models: a NuSMV-based approach', *International Journal of Software Engineering and Knowledge Engineering*, **20**(7), 987–1023, (2010).

[10] Antoni Ligęza, *Logical Foundations for Rule-Based Systems*, Springer-Verlag, Berlin, Heidelberg, 2006.

[11] Antoni Ligęza, 'BPMN – a logical model and property analysis', *Decision Making in Manufacturing and Services*, **5**(1-2), 57–67, (2011).

[12] Antoni Ligęza and Grzegorz J. Nalepa, 'Knowledge representation with granular attributive logic for XTT-based expert systems', in *FLAIRS-20: Proceedings of the 20th International Florida Artificial Intelligence Research Society Conference: Key West, Florida, May 7-9, 2007*, eds., David C. Wilson, Geoffrey C. J. Sutcliffe, and FLAIRS, pp. 530–535, Menlo Park, California, (may 2007). Florida Artificial Intelligence Research Society, AAAI Press.

[13] Antoni Ligęza and Grzegorz J. Nalepa, 'A study of methodological issues in design and development of rule-based systems: proposal of a new approach', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **1**(2), 117–137, (2011).

[14] Antoni Ligęza and Marcin Szpyrka, 'Reduction of tabular systems', in *Artificial Intelligence and Soft Computing - ICAISC 2004*, eds., Leszek Rutkowski, Jörg Siekmann, Ryszard Tadeusiewicz, and Lotfi Zadeh, volume 3070 of *Lecture Notes in Computer Science*, 903–908, Springer-Verlag (2004).

[15] Grzegorz Nalepa, Szymon Bobek, Antoni Ligęza, and Krzysztof Kaczor, 'Algorithms for rule inference in modularized rule bases', in *Rule-Based Reasoning, Programming, and Applications*, eds., Nick Bassiliades, Guido Governatori, and Adrian Paschke, volume 6826 of *Lecture Notes in Computer Science*, pp. 305–312. Springer-Verlag (2011).

[16] Grzegorz Nalepa, Szymon Bobek, Antoni Ligęza, and Krzysztof Kaczor, 'HalVA - rule analysis framework for XTT2 rules', in *Rule-Based Reasoning, Programming, and Applications*, eds., Nick Bassiliades, Guido Governatori, and Adrian Paschke, volume 6826 of *Lecture Notes in Computer Science*, pp. 337–344. Springer-Verlag (2011).

[17] Grzegorz J. Nalepa, *Semantic Knowledge Engineering. A Rule-Based Approach*, Wydawnictwa AGH, Kraków, 2011.

[18] Grzegorz J. Nalepa and Krzysztof Kluza, 'UML representation for rule-based application models with XTT2-based business rules', *International Journal of Software Engineering and Knowledge Engineering (IJSEKE)*, **22**(4), (2012). in press.

[19] Grzegorz J. Nalepa and Antoni Ligęza, 'HeKatE methodology, hybrid engineering of intelligent systems', *International Journal of Applied Mathematics and Computer Science*, **20**(1), 35–53, (March 2010).

[20] Michael Negnevitsky, *Artificial Intelligence. A Guide to Intelligent Systems*, Addison-Wesley, Harlow, England; London; New York, 2002. ISBN 0-201-71159-1.

[21] OMG, 'Production Rule Representation RFP', Technical report, Object Management Group, (2003).

[22] OMG, 'Semantics of Business Vocabulary and Business Rules (SBVR)', Technical Report dtc/06-03-02, Object Management Group, (2006).

[23] OMG, 'Business Process Model and Notation (BPMN): Version 2.0 specification', Technical Report formal/2011-01-03, Object Management Group, (January 2011).

[24] Chun Ouyang, Marlon Dumas Wil M.P. van der Aalst, and Arthur H.M. ter Hofstede, 'Translating BPMN to BPEL', Technical report, Faculty of Information Technology, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia Department of Technology Management, Eindhoven University of Technolog y, GPO Box 513, NL-5600 MB, The Netherlands, (2006).

[25] Marcin Szpyrka, Grzegorz J. Nalepa, Antoni Ligęza, and Krzysztof Kluza, 'Proposal of formal verification of selected BPMN models with Alvis modeling language', in *Intelligent Distributed Computing V. Proceedings of the 5th International Symposium on Intelligent Distributed Computing – IDC 2011, Delft, the Netherlands – October 2011*, eds., Frances M.T. Brazier, Kees Nieuwenhuis, Gregor Pavlin, Martijn Warnier, and Costin Badica, volume 382 of *Studies in Computational Intelligence*, 249–255, Springer-Verlag, (2011).

[26] M.T. Wynn, H.M.W. Verbeek, W.M.P. van der Aalst, A.H.M. ter Hofstede, and D. Edmond, 'Business process verification – finally a reality!', *Business Process Management Journal*, **1**(15), 74–92, (2009).

Web User Navigation Patterns Discovery as Knowledge Validation challenge

Paweł Weichbroth and **Mieczysław Owoc**[1]

Abstract. Internet resources as life environment of modern society with huge number of more or less conscious participants should be adjusted in terms of forms as well as content to users' needs. Knowledge acquired from web server logs in order to generate has to be validated. The aim of this paper is presentation of web usage mining as a quest for the knowledge validation process. A general concept of iterative and hybrid approach to discover user navigation patterns using web server logs is presented. Initial experiments on real website allow to define a new method of generated association rules refinement including specific knowledge validation techniques.

1 INTRODUCTION

Evolution of designing and developing Web sites from static to dynamic approach has enabled easy updates. Furthermore, intensive development and proliferation of WWW network resulted in other new modeling methods. It's obvious - being recognized and visited in the Web means that the content is up-to-date and satisfies its visitors. Widespread scope of content topics shared and presented on the Web site affects the size and depth level of its structures. This results in negative impression of presented content and weaker usability.

Usability of delivered information and knowledge depends on dynamically created user profiles as a result of Web mining and particularly user navigation patterns discovery process. Knowledge acquired from web server log files in order to generate navigation patterns embraces useful as well as non relevant association rules and has to be validated. Therefore the ultimate goal of this research is presentation of the method allowing for generated knowledge refinement.

These are very specific features of Web Mining procedures: temporal and massive data input, big differentiation of user types. They have direct impact on generated knowledge about website content and forms and in turn should be considered in Knowledge Validation (KV) framework. The paper is structured as follows. After presentation of related work, crucial definitions essential for this survey are proposed. Results of prepared experiments consisting of relationships between components are demonstrated in the next section. Crucial procedure for formulating knowledge discovery in such environment is investigated later. The paper ends with itemizing of conclusions and considering future research.

2 RELATED WORK

Recommendation systems help to address information overload by using discovered web users navigation patterns knowledge gained from web server log files. A problem for association rule recommendation systems (RS) is placed in dataset. It is often sparse because for any given user visit or object rank, it is difficult to find a sufficient number of common items in multiple user profiles. As a consequence, a RPS system has difficulty to generate recommendations, especially in collaborative filtering applications.

In [1] to solve above problem, some standard dimensionality reduction techniques were applied due to improved performance. This paper presents two different experiments. Sarwar et al. have explored one technology called Singular Value Decomposition (SVD) to reduce the dimensionality of recommender system databases. The second experiment compares the effectiveness of the two recommender systems at predicting *top-n* lists, based on a real-life customer purchase database from an *e*-Commerce site. Finally, results suggest that SVD can meet many of the challenges of recommender systems, under certain conditions.

Ad hoc exclusion of some potential useful items can be one of known deficiencies of this and other reduction of dimensions solutions hence they will not appear in the final results. Two solutions that address this problem were proposed by Fu et al. in [2]. The first solution assumes to rank all the discovered rules based on the degree of intersection between the left-hand side (antecedent) and active session of the user. Then, the *SurfLen* (client-server system) generates the top *k* recommendations. In addition to deal with sparse datasets - if users browsing histories intersect rarely, the system is unable to produce recommendations - the algorithm for ranking association rules was presented. The second solution takes advantage of collaborative filtering. The system is able to find "close neighbors" who represent similar interest to an active user. Then, based on that, a list of recommendations is generated.

Many collaborative filtering systems have few user ranks (opinions) compared to the large number of available documents. In this paper [3], Sarwar et al. define and implement a integration model for content-based ranks into a collaborative filtering. In addition, metrics for assessing the effectiveness filter bots and a system were identified and evaluated.

In this paper [4], Lin et al. a collaborative recommendation system based on association rules framework was proposed. The framework provides two measures for evaluating the association expressed by a rule: confidence and support. Moreover, the system generates association rules among users as well as among items.

3 DEFINITIONS

The web site space can be considered a universe U which consists of a sequential sets (P_i) where i=1...M. Each of sets P_i corresponds to unique user session where as each element of P_i is user's request for single page shared by a web site.

We consider only such subsets A of P_i ($A \subseteq P_i \subseteq U$) which appeared often enough. The frequency of subset A is defined as *support* (or support ratio) denoted as *support*(A)=$|\{i\ ;\ A \subseteq P_i\}|\ /\ M$. A "frequent" set is a set which support satisfies the minimum support value, denoted as *minsupport*. It is a user dependent value, often defined as cut-off as well.

We developed and applied an Apriori-like algorithm to mine frequent itemsets that is based on level-wise search. It scans a database recursively – a growing collections A are generated using smaller collections, especially the subsets of A of cardinality $|A|$-1,

[1] PHD Student, Katowice University of Economics, Katowice, Poland, pawel1739@gmail.com.
Department of Artificial Intelligence Systems, Wroclaw University of Economics, Wroclaw, Poland, email: mieczyslaw.owoc@ue.wroc.pl.

called hipersubsets. Formally, a set B is a hipersubset of a set A if and only if $\exists_{a\in A} A=B\cup\{a\}$.

An *association rule* is an implication of the form $A\rightarrow B$, where $B\subseteq P_i\subseteq U$ and $A \cap B = \emptyset$. The support of a rule is the percentage sessions in P that contains $A \cup B$. It can be seen as an estimate of the probability $Pr(A\cup B)$. The rule support is computed as follows: $support(A\rightarrow B) = support(A\cup B) / M$. The confidence of a rule $A\rightarrow B$ is the percentage of sessions in P that contain A also contain B. It can be seen as an estimate of the conditional probability $Pr(B\backslash A)$. The rule confidence is computed as follows: $confidence(A\rightarrow B) = support(A\cup B) / support(A)$. Confidence of the disjoint couple of sets $\{A,B\}$ can be read B under the condition A. If its value exceeds an arbitrarily determined level of minimum confidence (also defined by user), denoted as minconfidence, the pair $\{A,B\}$ will express an association rule $A\rightarrow B$.

Given a session data set D, the problem of mining association rules is to discover *relevant* (or *strong*) association rules in D which satisfy support and confidence greater than (or equal) to the user-specified minimum support and minimum confidence. Here, the keyword is “relevant” (or “strong”) which means, taking into account a user’s point of view, association rule mining process is complete.

Theorem 1: *Each subset of the frequent set is a frequent set.*

Proof 1: Let A be an arbitrary frequent set and $B\subseteq A$ be a subset of A. The set A is frequent, thus $support(A)=|\{i\ ;\ A\subseteq P_i\}|\ /\ M \geq$ minsupport. Since B is a subset of A, we have that $A\subseteq P_i \Rightarrow B\subseteq P_i$, thus $support(B) = |\{i\ ;\ B\subseteq P_i\}|\ /\ M \geq |\{i\ ;\ A\subseteq P_i\}|\ /\ M \geq minsupport$. Therefore the set B is also frequent which ends the proof. □

Theorem 2: *If $A\rightarrow B \cup C$ is a relevant association rule (A, B, C – pair wise distinct), then also $A\rightarrow B$, $A\rightarrow C$, $A\cup B\rightarrow C$ and $A\cup C\rightarrow B$ are relevant association rules.*

Proof 2: Let $A\rightarrow B \cup C$ be an user- relevant association rule. Then $minconfidence \leq confidence(A,B\cup C) = support(A\cup B\cup C)\ /\ support(A)$.

Let us consider $A\cup B \subseteq A\cup B\cup C$. Having in mind theorem 1, we get $support(A) \geq support(A\cup B) \geq support(A\cup B\cup C)$. As a result, $confidence(A,B) = support(A\cup B)\ /\ support(A) \geq support\ \{A\cup B\cup C\}\ /\ support(A) \geq minconfidence$. The set $\{A\cup B\}$ is frequent, so $A\rightarrow B$ is a relevant association rule. On the other hand, $confidence(A\cup B,C) = support(A\cup B\cup C)\ /support(A\cup B) \geq support\ \{A\cup B\cup C\}/support(A) \geq minconfidence$, because of a frequency of a set $A\cup B\cup C$, $A\cup B\rightarrow C$ is also a relevant association rule. □

Consequently, proofs for rules $A\rightarrow C$ and $A\cup C\rightarrow B$ are similar.

4 WEB USAGE MINING PROCESS

Methods and techniques commonly used in data mining are applied in Web Internet resources analysis. The main purpose of web usage mining is to discover users navigational patterns. Such knowledge can be the basics of recommendation systems in order to improve functionality (cross-selling) and usability (easier access) to Internet portals and services. First holistic and integrated (incl. different KDD steps) model of Web Usage Mining process was proposed by Cooley et al. in [5]. Afterwards, Srivastava et al. in [6] extended this model and apparently distinguished three phases: (*1*) preprocessing, (*2*) pattern discovery and (*3*) pattern analysis. This model has been widely applied – the references can be found e.g. in [7-14].

This model is complex and practice- oriented solution of solving problem of knowledge discovery from Web data repositories. The following particular tasks are subordinated to the mentioned phases. It allows for clear defining of user requirements from its point of view including data format and expected analysis results. However it seems to be reasonable to extend the model by data collection (see Fig. 1). During this phase data sources are defined based on data entry type and the same selection of necessary variables necessary variables including planned analysis are performed.

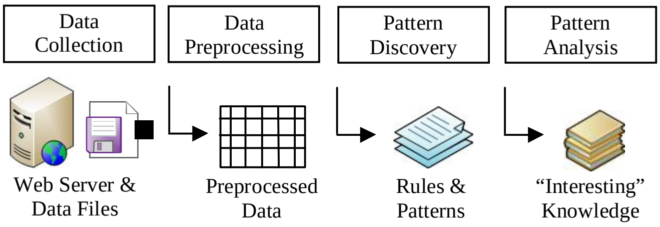


Figure 1. Extended Web Usage Mining process

Two dimensions can be defined in this model: e.g:

- phase – definition of data-oriented actions (tasks),
- object – definition of data types.

The whole process consists of the four phases; first three can be considered as machine-aided while the last one is human - expert oriented. In this frames, similarly to CRISP-DM specialized tasks are separate. We can distinguish four phases in web usage mining process:

1. Data collection, usage data can be collected from various sources like Web Server (e.g. log files), Client side (e.g. cookie files) or proxy servers.
2. Data preprocessing, this phase concerns raw data files which often includes web server log files. There can be distinguish following tasks: (*a*) data cleaning and filtering, (*b*) de-spidering, (*c*) user identification, (*d*) session identification and (*e*) path completion.
3. Pattern discovery, it can be seen as the data mining step when comparing to KDD. The following methods and techniques can be applied: (*a*) statistical techniques, (*b*) association rules, (*c*) sequential patterns, (*d*) clustering and (*e*) classification.
4. Pattern analysis, where domain expert evaluates discovered knowledge and optionally does performance assessment. If necessary, some modeling methods and techniques can be applied, such as clustering and classification.

In the next section, we refer to each phase briefly, emphasizing the most important issues which can be seen as a part of knowledge validation challenge.

5 EXPERIMENTS

In Web Usage Mining the major sources of data are Web Server and application server log files. We only used Web Server log files which are a set of Web users’ activity record, registered by onet.pl - one of the most recognized Web portals in Poland. Web Server log files contain full history of access requests to files, shared on the Web Server. Usually, http server vendors apply Common Log Format (CLF) to the log files associated with http protocol service. This format was developed by CERN and NCSA as the component of http protocol. According to this format, there are seven variables which were selected to be recorded. Complete description of CLF format and its variables can be found e.g. in [15].

5.1 Data collection

The data format of Web Server log files is specific in onet.pl (see Tab. 1) and includes seven variables. Sample entries, which represent Web users' activity, are presented below.

Table 1. The structure of raw Web Server log files.

	Variable					
Row	◊	♠	○	□	⌂	♣
1	140229	8654	2166	2	5723	724
2	140229	8654	2166	2	5723	725
3	140229	8655	2227	124	5086	8052
4	140229	8641	2310	26	1587	1007
5	140229	8654	2227	124	5086	8151

◊ time of the session, ♠ session ID, ○ user ID, □ service ID, ⌂ subservice ID, ♣ html file ID.

In our research, we used a web server log file which covers five hours of system activity - from 2PM to 7PM on 29th December 2008. We also want to notice the fact that dedicated solutions (e.g. scripts, services) on the Web Server side were implemented. They were intended to reduce globally useless data (e.g. spiders activity). In this case, we can determine that in the phase of data collection, some preprocessing tasks took(or have taken) place.

5.2 Data preprocessing

We used Linux operating system to operate on the raw data. To carry out this phase, a simple script in `awk` language was implemented to process the log files. Firstly, from six available variables just two were selected (♠ ♣) and separated from others. Secondly, these two were sorted accordingly to the session time (◊) and session identifier (♠). Thirdly, the sessions, where only one html page was requested by the users, were deleted from the log files. In such way, there were 2.132.581 unique sessions observed which would be taken into account in our experiments. The size of the log file was reduced from 4.018.853 KB to 512.934 KB. Again, it can be noticed that necessary tasks which typically had to be performed in the second phase were very limited. In addition, the source data was supplemented with text files which contained dictionaries corresponding to the URL name (e.g. 724 denotes to [www]). Finally, the processed file had the following format (session id, URL name):

```
8654 [www]
8654 [sport.html]
8654 [games.html]
```

5.3 Pattern discovery

In this phase, frequent Web users' access patterns are discovered from sessions. A few algorithms have been proposed to discover sequential patterns so far. During the literature study of KDD, we came across the algorithms such as AprioriAll [16], GSP [17], SPADE [18], PrefixSpan [19], ItemsetCode [20].

We have decided to use AprioriAll, although it is not the most efficient and recent one. It is suitable solution for our problem since we can make it more efficient by pruning most of the candidate sequences generated in each iteration. This can be done because of given constraints and requirements. We cannot assume that for every subsequent pair of pages in a sequence the former one must have a hyperlink to the latter one. The justification of that assumption is the fact that we do not have full access to the Web Server repository. Some files could have been deleted, moved to the archives or simply changed by the provider. On the other hand, the frequency of content update can be seen as very high - we have to keep in mind that approximately 7 million users per day visit this Web portal. So, it is obvious that the content hosted on the portal must be up-to-date. Furthermore and what seems to be the most important, the engine of recommendation system is able to generate links directly on web pages and does not require knowledge of inner-oriented structure and relationships of individual web pages.

Let $P = \{p_1, p_2, \dots , p_m\}$ be a set of web pages, called items. Database D represents a set of reconstructed sessions $S = \{s_1, s_2, \dots , s_n\}$. Each session s_n is a subset of P where $s_n \subseteq P$, called itemset.

We implemented Apriori-like algorithm (Algorithm 1) where input is a database of sessions D and output is a set of frequent sequences L.

Algorithm 1: AprioriAll

```
$L_1$ = find frequent 1-itemsets($D$);
$C_2$ = all possible combinations of $L_1 \boxtimes L_1$
$L_2$ = find frequent 2-itemsets($D$);
for ($k = 3$; $L_{k-1} = \emptyset$; $k$++) {
    $C_k$ = S_Apriori_gen($L_{k-1}$);
    for each transaction $S \in D$
        $C_t$ = subset($C_k$, $s$);
        for each candidate $c \in C_t$
            c.count++; }
    $L_k = \{c \in C_k \mid c.count \geq min\_sup\}$ }
return $L = \cup_k L_k$
```

procedure *S-apriori_gen*(L_{k-1} : frequent(k - 1) -itemsets

```
for each itemset $l_1 \in L_{k-1}$ and
for each itemset $l_2 \in L_{k-1}$
```

```
        if($(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge ... \wedge (l_1[k - 2] \neq l_2[k - 1])$)
          then {
              $c = 1_1 \boxtimes l_2$;
              if has_infrequent_subsequence($c$, $L_{k-1}$) then
                  delete $c$;
              else add $c$ to $C_k$; }
      return $C_k$;

procedure has_infrequent_subsequence($c$: candidate k- itemset);
$L_{k-1}$: frequent($k$ - 1)- itemsets;

      for each($k$ - 1)-subset $s \in c$
          if s $\notin L_{k-1}$ then
              return True else
              return False;
```

As a result the program returns frequent sequences with support for each. A simple example of a three element sequence is given below.

```
[www];[info]/science/item.html;[info]/world/item.html;0,02011
```

Interpretation of this particular sequence may be expressed in these words: "*Over 2 percent of anonymous users between 2PM and 7PM on 29th December 2008, first requested access to the home page, next opened science section then world section*".

Based on the set of frequent sequences L we are able to generate sequential association rules (SAR). In this scenario, a simple algorithm was developed (Algorithm 2). Let $R = \{ r_1, r_2, ... , r_m \}$be a set of SAR. In each rule antecedent (body of the rule, left side) and consequent (head of the rule, right side) must not be replaced. In other words, it guarantees the precise order of each element in the sequence. Also, a user is requested to provide the minimum confidence (*minconfidence*) threshold.

Algorithm 2: SAR generator

```
$R$={}
for all $I \in L$ do
$R = R \cup I \rightarrow \{\}$
$C_1 = \{\{i\} \mid i \in I\}$;
$k$:= 1;
while $C_k \neq \{\}$ do
  //extract all consequents of confident association rules
  $H_k := \{X \in C_k \mid confidence(X \Rightarrow I \setminus X, D) \geq min\_conf\}$
  //generate new candidate consequents
  for all $X, Y \in H_k$, $X[i] = Y[i]$ for $1 \leq i \leq k - 1$ do
    $I = X \cup \{Y[k]\}$
    if $\forall J \subset I, |J| = k : J \in H_k$ then
      $C_{k+1} = C_{k+1} \cup I$
    end if
  end for
  $k$++;
end while
//cumulate all association rules
$R := R \cup \{X \rightarrow I \setminus X \mid X \in H_1 \cup \cdots \cup H_k\}$
end for
```

As a result program returns a set of sequential association rules *R*. We give a simple example of such results, based on previously given frequent sequence.

```
r1:{www};{science.html}→{world.html};
0,02011; 0,964
r2:{www}→{science.html}{world.html};
0,02011; 0,652
r3:{www}→{science.html}; 0,0305; 0,00209
r4:{www}→{world.html}; 0,0294; 0,00152
```

Interpretation of the first rule may be expressed in these words: "*There is a 96,4% chance that a user who visited {www} and {science.html} pages, after them would also visit {world.html}*". In other words, the value of confidence ratio shows degree of rule reliability.

In this way we have shown that attractiveness of discovered knowledge is evaluated by two measures: support and confidence ratio. They are both specified by a user - it guarantees a definitive result.

5.4 Results. Pattern analysis

The aim of the research was to discover frequent sequences which represent web user navigation paths. In first iteration the level of support and confidence ration was respectively 0,01 and 0,9. Table 2 shows top ten (taking into account the highest support, denoted by percentage) one element frequent itemsets.

Table 2. Itemsets and its support.

Itemset	Support	Itemset	Support
www	78.48%	email/logout.html	14,09%
email/login.html	27.08%	email/folder/delete.html	10,80%
email/folder.html	25.49%	sport/football/news.html	10,22%
info/world/item.html	20.27%	sport/formula one/news.html	9,88%
email/folder/open.html	15,01%	info/homeland/item.html	9,18%

For instance, human interpretation of such knowledge might be expressed in these words: *“Over 27 percent of sessions include a page enabling user to log on the email service”*.

Entire volume of frequent sequences draws picture of the most popular content of the web portal. First conclusion which arises from this analysis to divide and group discovered knowledge to two different categories: (*1*) service- oriented and (*2*) content-oriented. First category relates to the hosted services via the web portal like email, advertisements, auctions, games, video on demand, dates. Second category relates to the shared content like information, news and plots. In this kind of the web portal, we are not able to recommend anything on the home page unless we violate user’s privacy. On the other hand, keeping in mind high level of usability and visibility, the objects’ arrangement should remain static. Therefore service- oriented content will not be considered to be recommended. Also, over 60% of discovered knowledge concerns users’ actions while using email service. As an example let us deliberate on sequence below:

```
[www];[email]/login.html;[email]/inbox.html;[
email]/new-message.html;[email]/logout.html
```

Such knowledge is useless for recommendation engine since it simply presents obvious and feasible actions during common session, restricted to single service. In this case, we decided to decrease the confidence ratio to four additional levels: 0,8; 0,7; 0,6 and 0,5. Table 3 shows the volume of discovered sequential association rules for five different levels of confidence ratio.

Table 3. The volume of SAR.

	Number of SAR				
	Confidence				
Number of items	◊	□	○	θ	⌂
2	65	79	95	118	142
3	169	197	254	349	430
4	196	225	315	480	585
5	114	132	209	328	407
6	28	32	68	111	146
total	**572**	**665**	**941**	**1386**	**1710**

minimum confidence: ◊ 0.9 □ 0.8 ○ 0.7 θ 0.6 ⌂ 0.5

It can be noticed easily that simple and apparent indication has occurred - on every lower level of confidence the number of rules have increased. Finally, for 0,5 confidence more than 1710 rules were discovered. These 1138 rules (difference between 0,5 and 0,9 confidence) are not likely to bring further knowledge of web data usage. Nevertheless, some interesting itemset groups were observed which will be added to knowledge base. An example of useful sequence is presented below:

```
[www];[sport/football/news.html];[sport/formu
la one/news.html];[info/homeland/item.html]
```

At this point, previously undiscovered knowledge shall be reviewed and compared to that, which is already stored in knowledge base. Moreover, if discovered knowledge was transferred to knowledge base and its resources are engaged by recommendation engine, we are able to track the process of knowledge validation. It means that two measures *recall* and *precision* will determine degree of knowledge adequacy. In other words, we are able to determine quality of recommendations produced by the system.

Another step, which is possible to undertake and promises to discover new rules, is to decrease the value of the minimum support. There is no general instruction with suggested value in any scenario. It is a subjective issue and relies on expert intuition and experience. In the next two iterations, we set up the value of minimum support respectively on 0,05 and 0,001. In our opinion it should be compromise between time-consumption and hardware requirements constraints.

Table 4. Frequent sequences for three different minimum support values.

	Frequent sequences				
	Quantity			Change (◊=100)	
Number of items	◊	□	○	□	○
1	64	103	276	61%	331%
2	167	307	1522	84%	811%
3	177	438	3279	147%	1753%
4	418	338	3768	-19%	801%
5	40	154	2625	285%	6463%
6	14	44	1193	214%	8421%
7	0	6	373	-	-
8	0	0	73	-	-
9	0	0	5	-	-
total	**880**	**1390**	**13114**	**58%**	**1390%**

minimum support: ◊ 0.01 □ 0.005 ○ 0.001

The program was executed under Eclipse environment with Java version 1.6.0 on IBM PC x86 computer with an Intel Core2 Quad processor 2.4 GHz and 2 GB Random Access Memory. Execution times are 55 minutes, 2 hours 30 minutes and 28 hours 16 minutes respectively to lower value of the minimum support. The determination of the factors influencing on the effectiveness of algorithm covers an remarkable research question. Unfortunately, we are not able to put forward straightforward answers.

Let us examine new sequences discovered throughout these two independent iterations. Though, we only focus on the set (○) because it is a superset of the other two sets (◊ and □). First of all, we can notice a enormous growth - almost 14 times larger set of frequent sequences was discovered, when value of minimum support was set up on 0,001 comparing to 0,01. This fact is worth deep consideration however our attention would be focused on the longest nine- items sequences which one of them is presented below:

```
[www]; [dates]/visitors.html;
       [dates]/email.html;
       [dates]/email/new.html;
       [dates]/index.html;
       [dates]/user-page/index.html;
       [dates]/user-page/index/k.html;
```

```
        [dates]/user-page/album.html;
        [dates]/album.html
```

We can definitely state that the results are unsatisfactory. The longest sequences present the interactions within one hosted service. The same situation can be observed taking into account eight- and seven- items sequences. Just six sequences in the latter one would possibly classified to be added to the knowledge base. These sequences present interesting navigation paths like:

```
[www]; [business]/pap.html
        [business]/news.html
        [info]/homeland/item.html
        [info]/world/item.html
        [business]/stock market/news.html
        [business]/company/market.html
```

We asked ourselves: was it necessary to decrease value of minimum support having in mind program time-consumption? Even some portion of discovered knowledge might be useful, there is a great risk that it would become inadequate. This situation happens when the content update is very often. Discovered sequences are not longer available to reach because its items (represented by links) are simply replaced by others.

In our approach we anticipated the last step which can help to discover relevant knowledge from web server log files. The assumption is simple: the domain expert shall determine service-oriented itemsets and any other irrelevant items and exclude them from the data. Then, the data is preprocessed again and process of web usage mining starts one more. Preliminary experiments have confirmed presented model in the next section.

6 VALIDATION MODEL

In this section, we present an iterative model for discovering and validating web user navigation patterns.

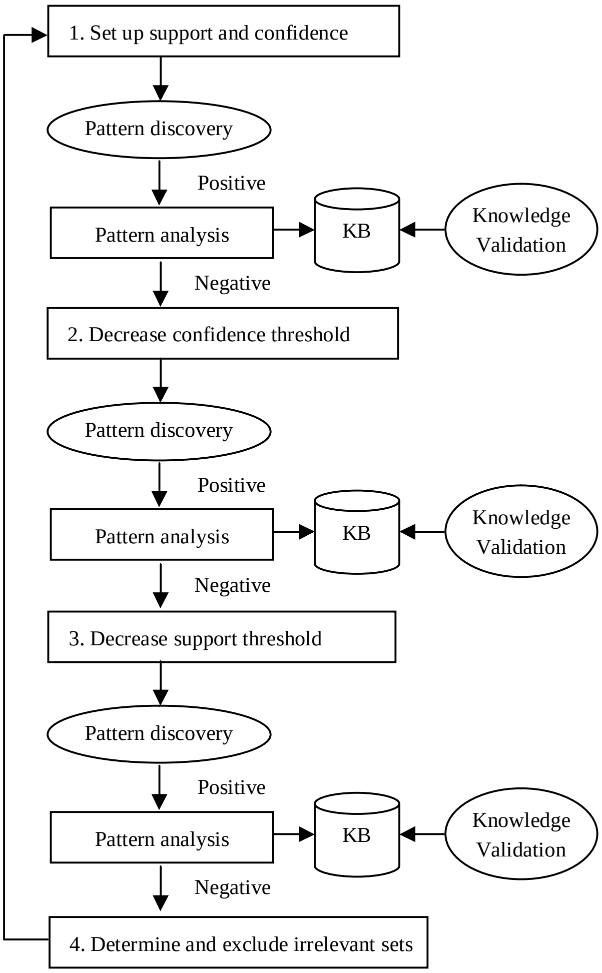


Figure 2. Iterative model for discovering and validating web user navigation patterns

In the first step, initial values for minimum support and confidence are set by a user. Next, the pattern discovery process is performed and thus analyzed and evaluated by a user. Positively assessed knowledge is put into the knowledge base from where inference engine produces a list of recommendations to active user sessions. The effectiveness of this process can be measured by two metrics: *recall* and *precision*. Therefore the knowledge base is validated by them, expressing the degree of its adequacy. If the values of metrics are not satisfied, in the second step the confidence value is decreased. Similarly, in the third step, a user can decrease the minimum support value. If even in this case, values of evaluation metrics are not satisfied, a user specifies additional constraint by excluding sets which are irrelevant. As a result, the process returns to its beginning.

7 CONCLUSIONS

In this paper, we introduced an interactive a user-driven model which addresses problem of validating web user navigation patterns. Experiments performed on web server file logs show that the model is useful and in some circumstances can be used in large-scale web site environments.

REFERENCES

[1] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, "*Application of Dimensionality Reduction in Recommender System - A Case Study*" Proc. ACM WEBKDD Workshop, ACM, (2000).

[2] X. Fu, J. Budzik and K.J. Hammond, "*Mining navigation history for recommendation*", ACM, 106-112, (2000).

[3] B.M. Sarwar, J.A. Konstan, A. Borchers, J. Herlocker, B. Miller and J. Riedl, "*Using filtering agents to improve prediction quality in the GroupLens research collaborative filtering system*", ACM, 345-354, (1998).

[4] W. Lin, S.A. Alvarez and C. Ruiz, "*Efficient Adaptive-Support Association Rule Mining for Recommender Systems*" Data Mining and Knowledge Discovery, **6**(1), 83-105, (2002)

[5] R. Cooley, B. Mobasher and J. Srivastava, "*Web Mining: Information and Pattern Discovery on the World Wide Web*". IEEE Computer Society, (1997).

[6] J. Srivastava, R. Cooley, M. Deshpande and P.N. Tan, "*Web usage mining: discovery and applications of usage patterns from web data*" ACM SIGKDD Explorations Newsletter, **1**(2), (2000).

[7] R. Kosala and H. Blockel, "*Web mining research: A survey*" Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mininig SIGKDD: GKDD Explorations, **1**, (2000).

[8] M. Eirinaki and M. Vazirgiannis, "*Web mining for web personalization*," ACM Trans. Internet Technology, **3**(1), 1-27, (2003).

[9] B. Mobasher, H. Dai, T. Luo and M. Nakagawa, "*Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization*" Data Mining and Knowledge Discovery, **6**(1), 61-82, (2002).

[10] D. Pierrakos, G. Paliouras, C. Papatheodorou and C.D. Spyropoulos, "*Web Usage Mining as a Tool for Personalization: A Survey*," User Modeling and User-Adapted Interaction, **13**(4), 311-372 (2003).

[11] F.M. Facca and P.L. Lanzi, "*Mining interesting knowledge from weblogs: a survey*" Data Knowledge Engineering, **53**(3), 225-241, (2005).

[12] M.A. Bayir, I.H. Toroslu, A. Cosar and G. Fidan, "*Smart Miner: a new framework for mining large scale web usage data*". ACM, 161-170, (2009).

[13] T. Staś, "*Wykorzystanie algorytmów mrowiskowych w procesie doskonalenia portali korporacyjnych*". PHD Thesis. Wydział Zarządzania. Katedra Inżynierii Wiedzy, Akademia Ekonomiczna im. Karola Adamieckiego w Katowicach, Katowice, (2008).

[14] P. Markellou, M. Rigou and S. Sirmakessis, "*Mining for Web Personalization*" Web Mininig: Applications and Techniques, A. Scime (ed.), Idea Group Reference, 27-49, (2005).

[15] W3C, "Logging Control In W3C httpd," http://www.w3.org/ Daemon/User/Config/Logging.html#common-logfile-format. (1995).

[16] R. Agrawal and R. Srikant, "*Mining Sequential Patterns*". IEEE Computer Society, pp. 3-14, (1995).

[17] R. Srikant and R. Agrawal, "*Mining Sequential Patterns: Generalizations and Performance Improvements*", Springer-Verlag, pp. 3-17, (1996).

[18] M.J. Zaki, "*SPADE: An Efficient Algorithm for Mining Frequent Sequences"*. Machine Learning, **42**(1/2), 31-60, (2001).

[19] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal and M.-C. Hsu, "*Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach*". IEEE Trans. on Knowledge and Data Engineering, **16**(11), 1424-1440, (2004).

[20] R. Ivancsy and I. Vajk, "*Frequent pattern mining in web log data*". Acta Polytechnica Hungarica, **3**(1), 77-90, (2006).

Kleenks: collaborative links in the Web of Data

Razvan Dinu[1] and **Andrei-Adnan Ismail**[2] and **Tiberiu Stratulat**[3] and **Jacques Ferber**[4]

Abstract. Linked Data is an initiative towards publishing and connecting structured data on the Web, creating what is called the Web of Data. This allows the creation of new types of applications, such as mash-ups and semantic searches, which make use of and integrate data coming from multiple online repositories. However, the large amount of content produced by blogs, wikis and social networks, which have become de facto standards for publishing content in Web 2.0, suggests that the growth of Web of Data could also be supported by adding a social, unstructured, collaborative dimension to it. In this paper we introduce "kleenks", which are collaborative links that combine structured and unstructured data by allowing users to add unstructured content to links, in addition to the RDF predicate. The quality and importance of such links can be evaluated by the community through classical mechanisms such as ratings and comments. This approach stimulates the emergence of more complex and abstract relations between entities, allowing people to take part in the Linked Data process and actively contribute to the growth of the Web of Data. We discuss how kleenks can be modeled on top of RDF and RDFS, making them easy to implement, and identify the main challenges to be addressed by a platform implementing the kleenks model. Finally, we introduce an online platform that successfully applies kleenks in the research domain by allowing researchers to create kleenks between articles, books and any other type of online media.

1 Introduction

1.1 Context

"This is what Linked Data is all about: it's about people doing their bit to produce a little bit, and it's all connecting.,, - Tim Berners-Lee, TED 2009.

Linked data is a movement trying to expose the world data in a structured format and to link it all together in meaningful ways. This concept has been gaining traction as more and more organizations are starting to expose their data in a structured, computer-understandable format, besides the traditional website. Until recent, the habit was this: if an organization owned some data and it wanted to expose it to the public, it created a website allowing users to explore it. However, it soon became obvious that this was not enough; humans were not the only ones interested in working with this data, sometimes even computers or software agents delegated by humans should be able to manipulate it. In the dawn of this era, the web crawling and screen scraping concepts appeared. Programs that contained specific parsing code for extracting knowledge out of raw HTML emerged, and they were named crawlers or scrapers. Due to the technical difficulties of doing NLP (Natural Language Processing), these programs would use the underlying regularities in the HTML structure to parse the structured data. Soon, a war broke out between content owners who did not want to expose their data to machines and humans aiding the machines in extracting the data by continuously adapting the parsers to changes in HTML structure and to security additions aimed at differentiating humans from crawlers.

In the center of this war comes Berners-Lee's concept of Linked Data. Linked data is no longer data exposed by machines for machines, but it is data exposed by humans for their fellow machines. The Linked Open Data (LOD) project is leading this movement of encouraging people to expose their data to machines in a meaningful format. Most of the projects put forward by LOD are projects in which humans are in the center of the process of generating linked data. Big names in the internet industry such as Facebook agree with this vision, as confirmed by the launch of Facebook Open Graph v2 initiative at the F8 conference in 2011[5]. This announcement is about making the transition from the singular "like" action that users could perform on the social platform to a multitude of actions, even custom ones definable by application developers: read, watch, listen, cook, try out, and so on. Given the large amount of data continuously generated by users on their social networks, this step will finally expose all that data internally as structured data.

The DBpedia project is a community effort to extract structured information from Wikipedia and to make this information accessible on the Web [2]. This is actually an attempt at automatically parsing the Wikipedia infoboxes (the boxes with highlighted information usually in the right part of the page) into RDF triples. This database of triples is maintained in permanent synchronization with Wikipedia by using a subscription to the live modifications feed. In this case, people still play a central role in the generation of data, but their actions have the creation of linked data only as an indirect consequence.

Freebase [3] is an initiative supported by Google to apply the wiki concept to the world's knowledge. A user interface and a RESTful API are provided to users in order to be able to collaboratively edit a database of triples spanning more than 125 million triples, linked by over 4000 types of links, from domains as diverse as science, arts & entertainment, sports or time and space. One of the main focuses of this project is the user-created ontology, which is constantly evolving and cannot possibly be a set of fixed existing ontologies, no matter how complete they are, due to user friendliness reasons. There is actually one interesting conclusion arising from this fact: using a distributed workforce to crowdsource structured data requires a compromise between data precision and data quantity.

1.2 Problem statement

As we have seen in the previous section, there is a growing need for exposing the world's data in a structured format, as confirmed

[1] University of Montpellier 2, LIRMM
[2] Politehnica University of Bucharest
[3] University of Montpellier 2, LIRMM
[4] University of Montpellier 2, LIRMM

[5] https://f8.facebook.com/

by industry giants and academia alike. There are a number of efforts trying to bridge this gap. Only to name a few:

- crowd-sourcing structured data from users; examples are Freebase and OpenStreetMap
- crowd-sourcing unstructured data from users, in a nearly-structured format; examples are Wikipedia and Facebook before the launching of Facebook Open Graph v2
- crawling / scraping data from unstructured data; this includes shopping, travel and news aggregators, just to give a few examples
- extracting entities and links from unstructured text using NLP (Natural Langauge Processing); one elocvent example of this is OpenCalais[6]

However, current efforts for structuring the web's data are mostly concentrated around describing entities and their properties, as shown in [2]. This is also the nature of the information usually found in web pages: in Wikipedia, each page is dedicated to one entity, and none to relations between entities. Also, most of the current approaches generate data through automated means, by parsing online data sources or exposing legacy databases in RDF format. This has two shortcomings: the only relations present in Linked Data are those detectable by a computer program (so only explicit relations can be detected), and also the decision of whether the data is correct or not is left to the computer. Moreover, the current quantity of available linked data in the largest such database was 4.7 billion RDF triples [2], compared to over 1 trillion of web pages in 2008 [7]. This tells us that the current approach of exposing the web's data in a structured form is not scalable enough when compared to the explosive growth of social content since the advent of Web 2.0 and the social web: tweets, statuses, blogs, wikis and forums are all very hard to understand for a computer program.

Therefore, it is our strongly held belief that general linked data would benefit from a social component, allowing its creation to be crowdsourced among enthusiasts, given that they are motivated correctly, without compromising data integrity. We envision that people should be able to easily create links between any two online entities identifiable by a unique URI and to associate extra information to these links. If this process of creating linked data is turned into a community process, the validity of the links can then be subjected to community scrutiny, a model that is not too scalable, but has proven to work given enough contributors in Wikipedia's case. The possibility of linking resources is already built in the HTML standard; however, the amount of extra information one can currently associate with a link is limited. Also, links in a webpage cannot generally be subjected to community examination for validity, and cannot be easily removed from the page.

A tool for editing links between entities and for visualising the most important links between contents is not available yet, and is a necessary step forward for communities to support the creation of linked data. However, this task of generating new linked data should be approached with care, since providing structured data requires a certain amount of rigor and time, whereas most people lack at least one of the two. This is why providing structured data for an ordinary user is still a challenge, as proved by the Freebase [8] project, and why currently linked data which is not generated automatically is created by experts or dedicated people.

1.3 Article outline

The remainder of the article is structured as follows. Section 2 presents relevant works that are related to the kleenk platform and how our platform relates to each of them. Section 3 introduces a scenario that will be used throughout the article to exemplify the utility of our proposed model and framework. Section 4 introduces the new type of link we propose, the kleenk, and discusses its formal definition and evaluation mechanisms. Section 5 presents major challenges that have to be overcome by an implementation of our proposed concepts. Section 6 discusses how kleenks can be modelled with existing theoretical frameworks, and why we have chosen RDF. In section 7 we present our current implementation of kleenks, a platform aimed at connecting scientific content through a crowd-sourcing mechanism. Finally, in the last section, we present our conclusions and future works.

2 Related works

Here, we have chosen a few relevant works that treat the same problems as mentioned previously: adding a social dimension to the web of data, using crowdsourcing to build up the web of data, or ways to open up linked data to the big public, which might be the only fighting chance of keeping up with the growth rate of online content.

ConceptWiki[9] tries to apply the wiki concept to linked data. It contains a list of concepts as specific as "an unit of thought". Any person with an account on the website can edit the concepts and there are two main sections on the website right now: WikiProteins (which contains information about proteins) and WikiPeople (which contains information about authors in the PubMed database). The WikiPeople sections seems to be populated by extracting information from PubMed, an important technique in order to encourage user adoption that our proof-of-concept implementation,
kleenk.com, currently misses. Simply put, users tend to consider a website more reliable if it has more content on it. However, for ConceptWiki, this content is entity-oriented and is created automatically by machines instead of being created by humans (just like in DBpedia). Users can edit the existing content or add a new one, but the quantity of information needed to complete the page of a person can be quite daunting, which is why we suspect that ConceptWiki isn't still adopted on a large scale. We have derived one very important lesson from this project: using machines in order to generate enough data to bootstrap a community is a very good idea, as long as it is not too complicated for humans to emulate what machines are doing (or said differently, machines do now know the difference between user friendliness and otherwise).

Last but not least, Facebook Open Graph (v2) is a recent development of the social networking giant, allowing people to publish their social life as something very similar to RDF triples. People can now connect themselves to other entities by verbs like watch, read and listen, instead of the traditional like. Friends can afterwards rate and comment these actions, therefore this approach has also a very strong community evaluation component. However, this platform lacks in two respects: the first is generality, as it only connects people with entities, and through a pretty limited amount of actions (Facebook has to approve all new actions, giving it complete control over the ontology of predicates that appear); the second is aggregated visualisation capabilities, which is actually what makes the web of data interesting for the regular user: the ability to discover new content by navigating from content to content.

[6] http://www.opencalais.com/

[7] http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

[8] http:/www.freebase.com

[9] http://conceptwiki.org/

The fact that there are a number of projects solving the same problem as us, some even approached by internet giants or the academia gives us the strength to believe that we are working on the right problem. However, our proposed solution is unique, in that it lets users easily create their own linked data, while giving them access to powerful visualisations, as we will shortly see in the next sections.

3 Working Scenario

We will use an academia-related working scenario in this article.

Rob is a PhD student in computer science and he is reading a lot of books and papers related to his subject, which is artificial intelligence. He is testing a lot of applications and algorithms to see how they perform in different scenarios. He would like to discuss his findings with other researchers to have their opinions and also make his results easily accessible. He is discussing with his friends and also he publishes multiple articles but he feels that the feedback is limited and delayed (at least a few months from an article submission to its publication). Rob also has some younger friends that study the same topic. Whenever they find a new interesting article or application they ask Rob about it: What's important about this article? How does this application relate to application Y? Rob could tell them to read his articles but that may take a lot more time and his friends may get confused and get lost in other information they might not need. He gives them the answer but he knows that there may be more students out there that would benefit from those answers. How can he structure this information, and where to put it, so that it can be easily found by all interested researchers?

4 Kleenks

In this section we will propose a solution to the problem stated in section 1.2. We start by considering a simplified model of the Web of Data which allows us to explain the role of our approach and how it fits in the existing landscape. We finish by identifying the main challenges for implementing our proposal.

4.1 Web of Data

We consider a simplified model for the Web of Data which consists of the following elements: *contents*, *entities*, *links*, *software agents*, *humans* and *ontologies*.

Contents represent any type of unstructured data such as text, images, sounds or videos and they may, or may not have, an URI that uniquely identifies them. Entities can represent anything such as places, people or articles and they are uniquely identified by URIs. Links connect two entities, have an associated type and they can represent any relation between entities. By software agent we understand any software application (desktop, web or mobile) that uses the Web of Data. Also, we consider that humans can access entities and links directly, making abstraction of the browser or any application in between. Finally, ontologies can be used by both humans and software agents to understand the links between entities.

4.2 A new perspective, a new type of links

Inspired by the explosion of content in Web 2.0, we believe that the Web of Data could also use an internal perspective in which links are created from the user's point of view. We believe that the Web of Data needs a new social, unstructured and collaborative dimension that would bring people, unstructured content, entities and links closer to each other (Figure 1).

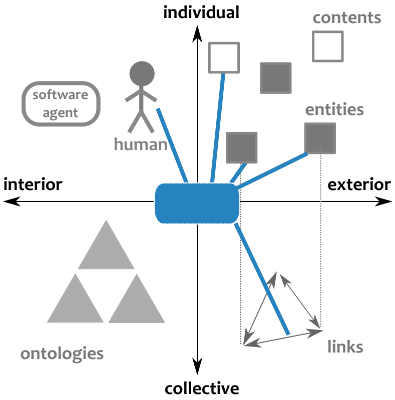


Figure 1. Social, Unstructured, Collaborative dimension to the Web of Data

We argue that this can be achieved through a new type of links, that we call *kleenks* (pronounced "clinks"), which are collaborative links created, evaluated and consumed by the users of the Web of Data. A kleenk (Figure 2) is a directed connection and consists of the following (below the words "entity", "content" and "link" have the meaning considered in the simplified model of the Web of Data from the beginning of this section):

1. **Source.** The source of a kleenk is an entity.
2. **Target.** The target of a kleenk is another entity.
3. **Type.** The type is a verb or expression that summarizes the link from the source to the target.
4. **Contents.** The contents represents the most important elements of a kleenk and they can have different roles:
   - *Description*. Descriptive contents can be simple text paragraphs, other media contents such as images and videos or even domain specific. They provide more details about the connection and they are added by the creator of a kleenk.
   - *Feedback*. As with descriptive contents, feedback contents can take any form but they are added by other participants to the kleenk (other people or software agents).
   - *Evaluation*. Evaluation contents must provide means to obtain quantitative data about the quality of a kleenk and they can take the form of ratings, like or thumb up/down buttons etc.

Kleenks are collaborative links because new content can be added to a kleenk at any time by its creator or by other participants. Kleenks have un unstructured dimension because the content added to a kleenk is in an unstructured form. Finally, a kleenk is social because it provides a mechanism for users to express their position (like, agree, disagree, etc.) with respect to it.

The term "kleenk" is actually a short version for collaborative link with a slightly different spelling since the term "clink" has been used in other works such as Project Xanadu[10] and we wanted to avoid confusion.

Let's take an example. Rob, from our first working scenario, reads a paper X that talks about an efficient implementation of an algorithm described in another paper Y. He will create a kleenk from the article

[10] http://www.xanadu.com

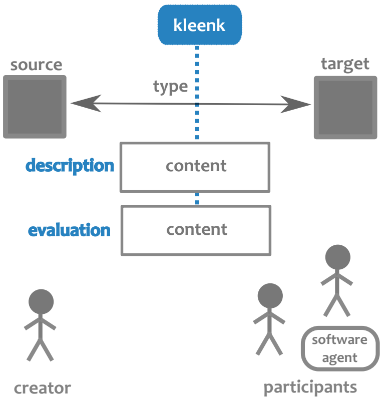


Figure 2. Elements of a kleenk

X to article Y with the type "efficient implementation of". Also, if the implementation is accessible on the internet he can also create a second kleenk from X to the implementation with the type "implemented here". As a description of the first kleenk he will provide a few details about what exactly makes the implementation efficient. Other researchers can express their opinion about the implementation directly on the kleenk, and comment for instance that the performance improvement is visible only on a particular class of input data. Other implementations can be kleenked to the same article X and the implementations can also be kleenked between them. Now, whenever an younger friend of Rob finds paper X he will quickly see the most important implementations of the algorithm and the relations with other important papers and they can continue their research without interruption.

4.3 Benefits and quality of kleenks

One main feature of kleenks is the ability to add unstructured content, in any form, to structured links. This has multiple benefits for both the user and the Web of Data. First because kleenks are richer in content than simple links, this makes them important on their own. Up until now, in the Web of Data, it is rare that links are very important on their own but rather in sets that describe an entity or a topic. We believe that making each link important on its own will engage people more in creating meaningful links.

Second, allowing people to create links with content will also facilitate the apparition of new links of high abstraction level that otherwise would have been impossible to extract automatically.

Allowing people to contribute to existing kleenks with new content is meant to make kleenks become more accurate and complete. However, as it has been seen in many projects such as Wikipedia and StackOverflow, an explicit evaluation system for user contributed content is necessary. The design of rating systems has been widely studied in computer science [7]. An overview of techniques that can be used to heuristically assess the relevance, quality and trustworthiness of data is given in [1].

Also, allowing social validation through mechanisms such as likes, agree/disagree or ratings allows important kleenks to step ahead of the less important ones guiding the users through what is important and what is less important. Of course the best way of validating a content can differ from domain to domain and each platforms that uses kleenks is free to choose the method that is more suitable.

5 Challenges

In the previous section we have introduced a new way of creating links in the Web of Data, at the conceptual level. This new type of links are called "kleenks" and they are collaborative links which contain unstructured content in addition to the typical RDF predicate. We believe that this approach will engage everyday users to take a more active part in building collaboratively the Web of Data and bring it to its full potential. However, implementing a system based on kleenks, be it targeted to a specific domain or as a general platform, raises a few challenges that must be properly addressed in order to be successfully used.

5.1 Access to entities

A kleenk, as an RDF triple, is a link that connects two entities and in addition it adds more content to the link. Letting regular users create such kleenks raises an important question: *"How will a user quickly select the entities he's interested in kleenking?"*.

The answer to this question depends on the type of platform: domain specific or general. In case of a domain specific platform it means that the user will kleenk entities he's working with. Usually these entities are already gathered in some databases and the kleenk platform only needs to integrate with these databases to provide quick search of the entities the user wants to kleenk.

On the other hand, a general platform is faced with a much more difficult question due to inherent ambiguities. If a user wants to use "Boston" as the source of a kleenk the platform has to decide whether it's about the city, the band or the basketball team. In this context we believe that semantic searches and large open databases such as DBpedia and Freebase will help in the disambiguation process.

Also, the user might want to kleenk things that don't yet have an URI and the platform must be able to create such URI's on the fly.

5.2 The ontologies for kleenks

Even though kleenks contain unstructured content, their type, as with RDF links, will still be a predicate in an ontology, allowing computers to have at least a basic understanding of what a kleenk means and use them in new ways. However, allowing users to create any type of links between entities means that it is very hard to develop a comprehensive ontology from the start. A kleenks platform would have to provide a mechanism that would allow users to define ontologies, such as in Freebase, or it must integrate with platforms that allow users to build ontologies such as MyOntology.

5.3 Visualization and privacy

Allowing users to create kleenks between any two entities has the potential of creating a very big number of kleenks. Users must be able to handle a big number of kleenks related to the entities that are of interest to them. Since kleenks form a graph structure, we can use visualisation techniques for graphs and create interactive ways of navigating the kleenks. We believe that since kleenks contain more content on the "edges" between the nodes, than just a simple predicate, more interactive and engaging visualizations can be built.

Since kleenks contain more content than simple RDF links and since most of this content will be based on the user's experience, the problem of the visibility of a kleenk must not be neglected. A user might want to create a kleenk between two entities and allow only a limited number of persons to see it. Also, kleenks can be used to collaboratively build some data (i.e. state of the art on a topic) which might, at least on its early stages, be visible only to a limited number of people. So, a kleenk platform must also provide proper mechanisms for kleenks' visibility.

6 Modeling kleenks

In this section we will look at the theoretical and technical aspects of modeling kleenks using existing techniques in semantic web. We will first analyze different alternatives and motivate our chose for one of them. Finally, we will give an example of what a kleenk might look like.

6.1 Theoretical model

Basically, the kleenk model could be seen as an extension of the RDF model with support for unstructured data. In the semantic web many extensions of the RDF model have been proposed during the last years. There are extensions dealing with temporal aspects [5], with imprecise information [8], provenance of data [4] or trust [9]. In [10] a general model based on annotation is proposed which generalizes most of the previous models.

All the above mentioned techniques are based on the named graph data model, a well known technique in semantic web to attach meta-information to a set of RDF triples. Even though these techniques could be applied to model kleenks, that would require that each kleenk has its own named graph (with its own URI), in order to associate the unstructured content with it.

A different technique, known under the name of RDF Reification, is described in the RDF specification [6]. This technique has well known limitations and weak points such as triple bloat and the fact that SPARQL queries need to be modified in order to work with reified statements. However, we believe that this techniques is the most suitable for modeling kleenks because a kleenk needs many different types of meta-information associated with it: creator, description content, feedback content (i.e. comments), evaluation content (i.e. ratings) and possibly other domain specific data.

7 kleenk.com

7.1 Description

kleenk.com [11] is an online collaborative platform for linking scientific content. The project's motto is: ”Smart-connecting scientific content”. It is allows users to link scientific contents, revealing other relations than citations, such as:

- paper P1 implements the algorithm in paper P2 (relation: ”implements algorithm in”)
- diagram D1 is an explanation for the theory in paper P2 (relation: ”explains the theory in”)
- algorithms A1 and A2 solve the same problem (relation: ”solves the same problem as”)

This kind of relation is not easy to extract neither by an automated program, and nor by humans that are just starting their research in a certain area. In Europe, the first year of a PhD program is usually dedicated to researching the state of the art, which consists of reading many scientific contributions by other authors and creating mental links like those mentioned previously. Given the exploding number of scientific works, conferences and journals it is hard to keep up-to-date even for a scientific advisor, which makes the work of a starting researcher even harder. Kleenk actually solves this problem by allowing the community to create and visualise kleenks between the contents.

This platform is aimed at the following groups of persons:

- PhD students which need community guidance in order to read the most relevant and up-to-date materials related to their subject
- professional researchers who need to stay in touch with the vibrant scientific community's developments
- other people interested in quickly gaining an overview of a scientific domain

The platform allows the easy selection of content to kleenk from a number of sources by manually adding it, importing it from web pages (such as ACM or IEEE public pages of articles) and even by importing BibTeX bibliography files. Once all the content a user wants to kleenk is available in the platform, the user can start creating kleenks by selecting a source and destination content.

After they are created, kleenks can be shared with research fellows or made public, and grouped around meaningful ideas using tags. Every time a new content is created or updated, the interested users are notified using their personal news feed. Therefore, changes to a kleenk or any comment reach out across the entire community instantly.

Authors have the chance to kleenk their own papers to existing ones, and by subjecting these kleenks to the community scrutiny, the platform makes it possible for them to obtain early feedback for their ideas. In today's society, when the internet allows information to be propagated from one end of the world to another in seconds, the traditional peer review system is becoming more and more criticized due to the number of months passed from submitting the work to actual post-publication feedback from the scientific community. Our service aims to complement the quality and thoroughness of the peer review system with the opinion of the crowd. One important observation is that the opinion of the crowd is not necessarily misinformed, as proven lately by the tremendously successful service for programmers StackOverflow [12]. This website is a collaborative question answering system, with world renown experts easily connecting and answering each others' questions. We think that the scientific community would benefit from a low-latency alternative to obtaining feedback for a piece of work.

7.2 Implementation of the theoretical framework

Having earlier detailed the kleenk model and characteristics, we will now underline which instantiation of the general principles was used in order to implement this knowledge sharing platform. First of all, in our particular case, the kleenk has the following elements:

- **the source, destination and type** - these are also present in the general model
- **the description** - this is specific to this pair of content, and represents a more detailed explanation of the type. It should be used in order to motivate the choice of type and to give more relevant results
- **comments** - since each kleenk has its own set of comments, these can be used in order to discuss the relevance of the link and to give extra information by anyone who can see it. These are similar to

[11] http://app.kleenk.com

[12] http://www.stackoverflow.com

Wikipedia's talk pages, which are used by contributors to clarify informations in the main page

- **ratings** - together with ratings, these allow the community to evaluate the quality of a kleenk. In the visualisation, kleenks with better community score (which is computed from the ratings, number of comments, number of views and a number of other metrics) are displayed with a thicker connecting line, signifying a greater importance. Ideally, an user who is interesting in exploring the web of scientific articles will first navigate the most important kleenks.
- **privacy level** - as already mentioned in the general model, there should be a privacy setting associated with each kleenk. This allows users to first try out their own ideas in a personal incubator before promoting them to the whole community. In our implementation, there are 3 privacy levels: private (visible only to the owner), public (visible to anyone) and shared (visible to research fellows, which can be added through a dedicated page, given that they also agree).
- **tags** - each kleenk can be part of one or more tags. This is actually a mechanism for grouping tags related to the same idea or topic under a single name. For example, when writing this article, the authors created a "Kleenk Article" tag which contained the relevant bibliographic items and the kleenks between them.

The visualisation of the graph induced by the kleenks is done, as mentioned in the description of the general model, using consacrated layout methods. Specifically, in our case, we use an attraction-force model.

kleenk.com is a linked data application, conforming to Berner-Lee's vision of the future of the web. Contents, kleenks and tags all have persistent URIs that can be dereferenced in order to obtain linked data. One other interesting side-effect of this is that interesting scientific applications can emerge on top on the data contributed by the users to kleenk. For example, new scientometric indicators based on kleenks could be computed by a 3^{rd} party application.

7.3 Use case example

7.3.1 *Obtaining feedback for a recently published article*

Alice is a fresh PhD student in Semantic Web, who is overwhelmed by the vast amount of publications on this topic. Being a first year student, she has to complete a document describing the state of the art by the end of the year. Being a Facebook user, it's easy for her to create an account using one click on kleenk.com, since it features integration with Facebook's login service. Once logged in, she adds her colleagues who already have a Kleenk account as research fellows and now can easily see their shared tags. She studies the visualisations and grows to see a few important articles which are in the center of most tags, and starts reading them. Since she pays close attention to her news feed, she can easily see in real time what connections her colleagues are creating, and they all obtain quick feedback from their advisor, via comments and ratings.

Since she will be writing a survey article as well, she started creating a tag specifically for the bibliography of the article. First, the tag is private, since it is a work in progress and she doesn't want to share it with anyone. As the text of the article and the bibliography mature, she changes the visibility of the tag from private to shared, so that her research fellows can express their opinion on the connections she is making. After receiving the final approval for publication, she makes the tag public and includes the visualisation of the bibliography in a presentation for her department.

8 Conclusions and Future Works

This article discusses the current context of the Web of Data, analyzes a few of its current limitations and focuses on the need to engage regular users in the creation of semantic links. We propose a new approach inspired by the success of Web 2.0 techniques such as wikis, blogs and social networks.

The main contribution of this paper is the concept of *kleenk* which is a collaborative link that contains unstructured data in addition to the classical RDF predicate. We discuss the importance of allowing users to add unstructured data to the Web of Data and how this approach could lead to the creation of links which would otherwise be impossible to automatically parse from existing datasets.

We also identify the main challenges of a platform allowing users to create kleenks: access to entities, collaborative ontology creation, visualization of kleenks and privacy. These challenges have to be properly addressed for a system to succeed in applying kleenks. We finish by introducing a free online platform, www.kleenk.com, which applies successfully the concept of kleenk in the scientific research domain and discuss how the identified challenges have been addressed.

Future works include:

- testing kleenks in other domains in order to see what would be the specific problems in adopting them for those domains
- building a common kleenk schema in order to describe kleenks
- defining scientometric metrics which are kleenk-related instead of the old citation-related approachess
- populating the kleenk.com database automatically with kleenks for citations in order to bootstrap the community use

REFERENCES

[1] Christian Bizer and Richard Cyganiak. Quality-driven information filtering using the WIQA policy framework. *Web Semant.*, 7:1–10, January 2009.

[2] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, September 2009.

[3] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM, 2008.

[4] Renata Dividino, Sergej Sizov, Steffen Staab, and Bernhard Schueler. Querying for provenance, trust, uncertainty and other meta knowledge in RDF. *Web Semant.*, 7:204–219, September 2009.

[5] Claudio Gutierrez, Carlos A. Hurtado, and Alejandro Vaisman. Introducing Time into RDF. *IEEE Trans. on Knowl. and Data Eng.*, 19:207–218, February 2007.

[6] http://www.w3.org/TR/rdf primer/. RDF Primer, February 2004.

[7] Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of Trust and Reputation Systems for Online Service Provision. *Decis. Support Syst.*, 43:618–644, March 2007.

[8] Mauro Mazzieri and Aldo Franco Dragoni. Uncertainty reasoning for the semantic web i. chapter A Fuzzy Semantics for the Resource Description Framework, pages 244–261. Springer-Verlag, Berlin, Heidelberg, 2008.

[9] Simon Schenk. On the Semantics of Trust and Caching in the Semantic Web. In *Proceedings of the 7th International Conference on The Semantic Web*, ISWC '08, pages 533–549, Berlin, Heidelberg, 2008. Springer-Verlag.

[10] Antoine Zimmermann, Nuno Lopes, Axel Polleres, and Umberto Straccia. A General Framework for Representing, Reasoning and Querying with Annotated Semantic Web Data. *Elements*, pages 1437–1442, 2011.

From Knowledge transmission to Sign sharing: Semiotic Web as a new paradigm for Teaching and Learning in the Future Internet

Noël Conruyt[1] and Véronique Sébastien[1] and Olivier Sébastien[1] and David Grosser[1] and Didier Sébastien[1]

Abstract. In the 21st century, with the advent of ultra high-speed broadband networks (1Gb per second), the Internet will offer new opportunities for innovators to design qualitative services and applications. Indeed, the challenge of such e-services is not only on the technological aspects of Internet with new infrastructures and architectures to conceive. The reality is also on its human and multimedia content delivery, with innovative philosophies of communication to apply in this digital and virtual age. In the context of Teaching and Learning as a human-centered design approach, we propose a new paradigm for thinking the Web, called the Web of Signs, rather than the Web of things. It focuses on the process of making knowledge by sharing signs and significations (Semiotic Web), more than on knowledge transmission with intelligent object representations (Semantic Web). Sign management is the shift of paradigm for education with ICT (e-Education) that we have investigated in such domains as enhancing natural and cultural heritage. In this paper, we will present this concept and illustrate it with two examples issued from La Reunion Island projects in instrumental e-Learning (@-MUSE) and biodiversity informatics (IKBS). This Sign management method was experimented in the frame of our Living Lab in Teaching and Learning at University of Reunion Island.

1 INTRODUCTION

The Future of Internet is not only a matter of technological, economical, or societal awareness; it is also grounded in individual, environmental and cultural values. Psychological, ethical, biological and emotional properties are indeed drivers of the Future Internet in a perspective of sustainable development of services with people. Although the Internet is the interconnection of networks of computers, it delivers interactive human-machine services such as the Web or Email [1]. The Web is an information service available on Internet with access to personalized documents, images and other resources interrelated together by hyperlinks and referenced with Uniform Resource Identifiers (URIs). Email is also a communication service available on the Internet. Nevertheless, Information and Communication Technologies (ICT) are not only oriented on technologies, but also convey human contents (data, information, and knowledge) that are communicated between end-users. At this upper level, co-designing e-services are means to connect producers and consumers of multimedia contents, in order that infrastructures of Future Internet meet user needs [2]. These principles have been adopted since 2006 by the European Network of Living Labs and are developed in the frame of corresponding literature [3].

Our idea is that in the Future Internet, we must not only pay attention to the quantity of information that is exchanged at higher speed between internauts (which is a techno-centric and economic perspective), but also to the quality of information communicated between people for them to be educated and aware of the richness and fragility of their environment. In the first case, users become stunned prosumers (producers and consumers of information), and in the second case they are simply responsible citizens (sensitive and educated people) in a closed world that must be preserved for the next generations.

In order to deliver such a holistic e-service in education, we introduce our methodology of Sign management in the first part of this paper. Then we explain how we organize the different types of Web that are part of the Semiotic Web, what makes its sense, and how to pass from Knowledge transmission to Sign sharing. This is illustrated with two examples taken from ICT projects for music education and biodiversity management. The conclusion emphasizes the need for repositioning human concerns at the center of technologies, and why we should favor the development of Living Labs philosophies.

2 SIGN MANAGEMENT

The reality of Future Internet is that it supports both technological and content services over the physical network. But in the 21st century, the technology must be at the service of human content and not the contrary. Indeed with Web 2.0, we have entered an era where usage is the rule for making e-services. Personalization of product/services accessible throughout the Internet is becoming more and more important as innovation is opening [4] and democratizing [5]. But we will have also to manage the quality of information that is exchanged between people in order that knowledgeable persons can express their know-how and be acknowledged [6] for it. In the context of climate change, biodiversity loss, pollution and globalization, it is urgent that the Future of Internet enhances scientific voices at human level.

But this endeavor cannot be led only by managing knowledge of specialists with the technology of Semantic Web, so-called Web 3.0 [7]. Knowledge management is not enough for the Future Internet. Firstly, knowledge cannot be managed because it resides between the ears of somebody (tacit knowledge). Only information that is transmitted between persons can be managed. Secondly, Knowledge can be found in books written by specialists (explicit

[1] LIM-IREMIA, EA 25-25, University of Reunion Island, 97490, Sainte-Clotilde, France, email: noel.conruyt@univ-reunion.fr

knowledge), but this is dead knowledge that cannot be updated. Knowledge is the result of a long experience of some experts that have experimented a lot of cases in the fields, and formed their know-how by compiling them in their mind. This know-how is living knowledge and it can be managed with multimedia contents (see below for music learning). Thirdly, the response of ICT to tackle knowledge management is to propose Semantic Web as a solution. Indeed, this is necessary in the context of representing objects of knowledge in computers with formats coming from description logics (RDF, OWL), but it is not sufficient as far as this technology cannot capture the signification of these objects for different individuals, i.e. Subjects.

In the context of enhancing Knowledge with ICT, we propose Sign Management as a shift of paradigm for the Future Internet. It emphasizes the engineering and use of data, information and knowledge from the viewpoint of a Subject. This concept is derived from the pragmatic Peirce's theory of semiotics with a Sign's correspondence of the Subject to its Object. From this philosophical viewpoint, a Sign, or representamen, is something that stands to somebody for something in some respect or capacity [8]. From our computer science analysis, Data (Object) is the content of the Sign (something), Information, a multi-layered concept with Latin roots ('informatio' = to give a form) is its form, and Knowledge is its sense or meaning, i.e. no-thing. The notion of Sign is then more central than knowledge for our purpose of designing e-services.

In Figure 1, we define a Sign as the interpretation of an Object by a Subject at a given time and place, which takes into account its content (Data, facts, events), its form (Information), and its sense or meaning (Knowledge).

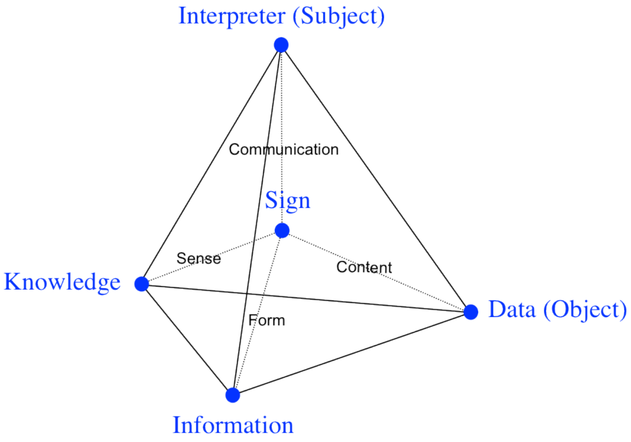


Figure 1. The tetrahedron of the Sign

Then, we introduce Sign-ification, the continuous process of using Signs in human thinking for acquiring Objects interpreted by Subjects. This signification process or Semiosis takes the different components of the Sign in a certain order to make a decision: first comes the Subject or Interpreter who is receptive to his milieu or "Umwelt" [9], and who cares about Information to act in a certain direction (volition), then occurs the searched Data (Object) to position himself in space and time (action), then Knowledge is activated in his memory to compare the actual situation with his past experiences and make an hypothesis for taking a decision (cognition). The Signification or the building of the sign communicates the process iteratively in a reflexive way (memorize new knowledge) or communicates the result (interpretation) as information to his environment (exteriorization), see Figure 2.

Semiosis is similar to the working principle of inference engine that was modeled in expert systems: the evaluation-execution cycle [10]. The difference is that Signification integrates the Subject in the process, and this integration is therefore more meaningful to humans than to machines. The Subject operates on Signs in two phases: reflection and action. These phases are linked in a reflexive cycle with a semiotic spiral shape including six moments: 1) to desire, 2) to do, 3) to know, 4) to interpret, 5) to know-how for oneself, 6) to communicate to others (Figure 2). The semiosis spiral is included in the tetrahedron of the Sign.

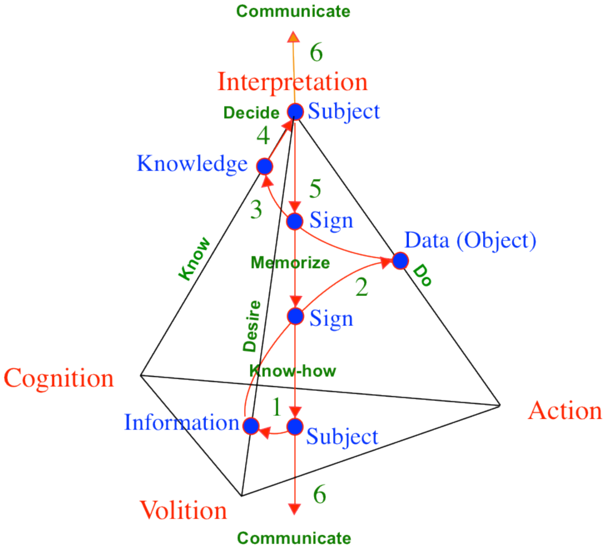


Figure 2. The signification process for Sign management

Consequently, Signification is the key psychological process that makes sense for practising usage based research and development with people by communicating data, information and knowledge. Signification is the kernel of Semiotic Web although Representation is at the root of Semantic Web. Both are necessary to co-design e-services in the Future Internet, but from our experience, don't miss Sign management and Semiotic Web if you want to co-design e-services with end-users!

3 SEMIOTIC WEB

Making sense or signifying is a biological characteristic that cannot be eluded in the Future Internet. We are acting now on a limited planet and the objective is to render services to human beings and become responsible rather than serve oneself and consume even more energy and matter with the help of computers.

When an organism or an individual seeks for something, his attitude is to pay attention to events of his environment that go in the sense (direction) of what he searches. The primary intention of a microorganism such as bacteria is "good sense": it wants to capture information from the milieu to develop itself and stay alive [11]. Human development follows the same schema of self-organized living systems at more complex levels than these physiological and safety needs. They are those that have been defined in the hierarchy of fundamental individual needs: love, belonging, esteem, self actualization [12]. As a consequence, we hold that before being able to make "true sense", i.e. adopt a scientific rationale, the objective of individuals is to respond to psychological needs (desire, pleasure, identity, etc.). This theory of human motivation is a natural and cultural hypothesis, which is corroborated by Umwelt [9], Activity [13] and Semiotic [8] pragmatic theories. These life and logical sciences are components of the Biosemiotics interdisciplinary research [14], which was introduced before the advent of Internet as the "Semiotic Web" [15].

Semantic Web is the dyadic combination of form and sense of the linguistic Sign [16], taken as a signifier (form) and signified (sense). It is rational. Semiotic Web is more generic and living. It complements the Semantic Web (form and sense) with the referents (content) that are observed data (interpretations) geo-referenced in a 3D information world (Immersive Web) as Web Services by subjects pertaining to communities of practice (Social Web 2.0). This makes our Sign management ecosystem a tetrahedron model (Figure 3) that is more involved in concrete life with end-users on a specific territory such as Reunion Island.

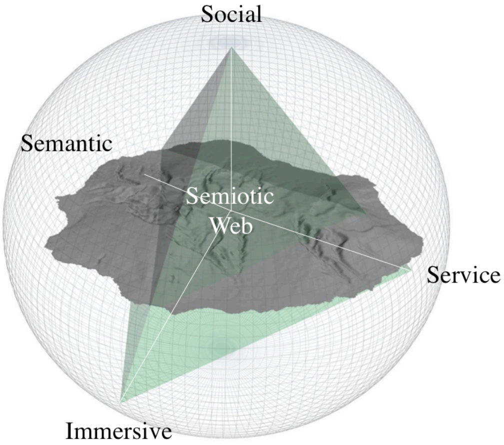


Figure 3. The Situated Service, Social, Semantic and Immersive Web

The Web of Signs combines:

1. The Web of Data and Objects, i.e. the flow of raw and digital contents produced by specialists (teachers) and transmitted by engineers in databases and knowledge bases in the frame of an Information System (one-way flow), but progressively becoming interoperable through Web services with other Information Systems,
2. The Web of Subjects, i.e. a bidirectional communication platform between users (teachers and learners) using different e-services within a community of practice to exchange interpretations of data and objects, and negotiate their value,
3. The Web of Information that is geo localized in attractive virtual worlds representing the real landscape (metaverses), and accessible at any time, anywhere, on any devices (mobiquity).
4. The Web of Knowledge for machines to communicate logically on the basis of a formal, open and semantic representation of data and objects,

At the University of Reunion Island, we have investigated each of these dimensions that are converging to form what we call the Semiotic Web. As the World is an Island and as Reunion Island is a small world, we designed our Living Lab as a small laboratory for Teaching and Learning Sciences and Arts by Playing [17]. Indeed, edutainment is one of the pillars of the Future Internet [18]. With game-based learning, we consider that we can play seriously to better know our environment and then better protect it.

For biodiversity management for example, we co-designed an Immersive Biodiversity Information Service (IBIS) for helping biologists and amateurs to access to forest and coral reef species information. This Teaching and Learning tool intends to use different modules dedicated to certain functionalities at different levels of data, information and knowledge and let them communicate by using Web Services [19].

In the spirit of Web 2.0 technologies, we participate to the ViBRANT FP7 project [20] that uses the Scratchpads for data sharing. Using a content management system (Drupal), Scratchpads 2.0 enables bottom up, collaborative work between all types of naturalists, from researchers to amateurs. This Social Web tool supports communities of taxonomists to build, share, manage and publish their data in open access.

For computer-aided taxonomy, we developed an Iterative Knowledge Base System platform called IKBS [21] with some taxonomists. It is based on a knowledge acquisition method and an observing guide for describing biological objects, i.e. the descriptive logics in life Sciences [22]. Our descriptive logics must not be confused with description logics (RDF, OWL) of the Semantic Web because they are the rules of thumb of experts for making descriptive models (ontologies) and describing cases. The objective of this Research tool in Biodiversity Informatics is to help biologists classify and identify a specimen correctly from an expert viewpoint by using onto-terminologies (ontologies + thesaurus).

4 FROM KNOWLEDGE TRANSMISSION TO SIGN SHARING

Knowledge is subjective in the paradigm of Sign management: it cannot be taken for granted without putting it into use, mediated and negotiated with other Subjects on a meeting place, which we called a Creativity Platform [23]. What can be managed is called descriptive or declarative knowledge: it is the communication of justified true beliefs propositions from one Subject made explicit. The formal interpretation process from observation to hypotheses, conjectures and rules is called signification of knowledge on the human communication side of the Sign. It is called representation or codification of knowledge on the machine information side of the Sign. Apart from being described, this interpretation process can be shown with artifacts to illustrate the description ("draw me a sheep", says the little prince!). Sign management wants to enhance this aspect of multimedia illustration of interpretations to facilitate transmission and sharing of knowledge through the communication of the Subject (see the fourth communication part of the sign in Figure 1).

In knowledge management, propositional knowledge is taken mostly in the sense of scientific knowledge, considered as objective in scientific books, and providing the know-that or know-what. Ryle in [24] has shown that this is confusing. In the sense of subjective knowledge taken as "I know that or I know what", there is the other sort of knowledge called know-how. It is "the knowledge of how to do things", i.e. what the subjects can show through their interpretations when they practice their activity (there is a difference between the recipe and the cooking of the recipe, isn't it?). And some people do the activity better than others. They are called the experts. As such, know-how is closer to data (Praxis) and information (Techne) than to knowledge (Scientia). Finally, know-how and know-that or know-what are different categories of knowledge and should not be conflated [25]. Knowledge synthesizes what makes sense in the head of skilled persons for doing well the tasks of their activity.

Starting from these differences of interpretations about the term of knowledge, and considering the domain of activity that we want to deal with, i.e. education with ICT, we prefer to focus on managing interpretations, and firstly the good ones from professors. Sign management manages live knowledge, i.e. subjective objects found in interpretations of real subjects on the scene (live performances) rather than objective entities found in publications (bookish knowledge).

In this context of managing know-how rather than knowledge, we have set-up our Living Lab in Reunion Island on the thematic of Teaching and Learning by Playing [26]. The sharing of expertise with ICT is our added value in education for some specific domains such as managing biodiversity, performing art (music, dance, etc.), speaking a language, welcoming tourists, or cooking. These niches can be enhanced with ICT in a sustainable manner by following some innovative methods. For example, sign or know-how management produces sign bases that are made of interpretations for knowing how-to-do things with multimedia content and not only knowing what are these things in textual Knowledge bases.

Finally, a Sign is a semiotic and dynamic Object issued from a Subject and composed of four parts, Data, Information, Knowledge and Communication. Our Sign paradigm uses a fouradic representation (a regular tetrahedron, see Figures 1 and 2) instead of the triadic sign representation that lets the Subject outside of the Semiosis process. All these subjective components communicate together to build a chain of significations and representations that we want to capture.

Sign management makes explicit the subjective view of doing arts and sciences. Our aim is to compare different interpretations of subjects about objects through transmitting and sharing them on a physical and virtual space dedicated to a special type of e-service, i.e. in instrumental e-learning or biodiversity informatics (see below). For the purpose of co-designing such a service with ICT, the Creativity Platform is the co-working, learning and communication space for researchers and developers, businesses and users, aimed at collectively defining the characteristics of e-services in order to ensure the most direct correspondence between expectations and use [27].

5 SIGN SHARING IN MUSIC TEACHING AND LEARNING

Sharing Signs is particularly relevant in artistic fields, where a perfect synchronization between gestures, senses and feelings is essential in order to produce original and beautiful works.

In this frame, the @-MUSE project (@nnotation platform for MUSical Education) aims at constituting a Musical Sign Base (MSB) with the interactions coming from a community of musicians. This project benefits from the experience we accumulated in the field of instrumental e-Learning in Reunion Island, from various mock-ups to complete projects such as e-Guitare [23]. Figure 4 sums up our research process in this domain, based on a Creativity Platform.

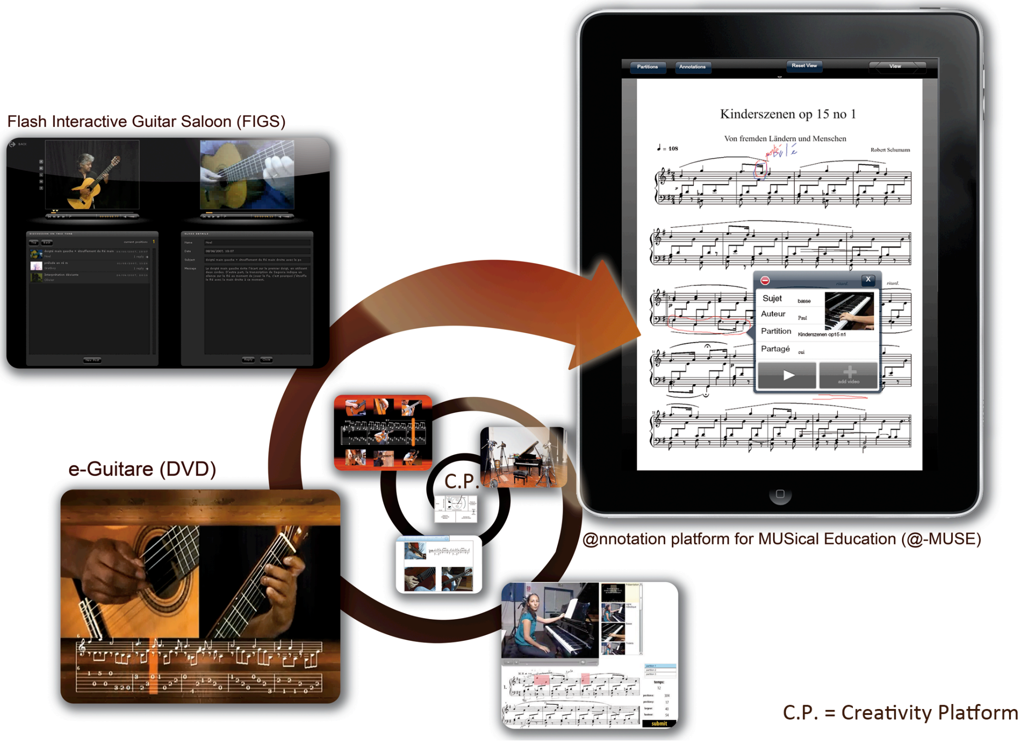

Flash Interactive Guitar Saloon (FIGS)
e-Guitare (DVD)
C.P.
Kinderszenen op 15 no 1
Von fremden Ländern und Menschen
Robert Schumann
Sujet
Auteur
Partition
Partagé
@nnotation platform for MUSical Education (@-MUSE)
C.P. = Creativity Platform

Figure 4. Instrumental e-Learning services co-designed on a Creativity Platform

While the different versions of e-Guitare were more centered on the teacher performance, the FIGS (Flash Interactive Guitar Saloon) service was more axed on the dialog between learners and teachers through an online glosses system. What principally emerged from these projects was the need to facilitate the creation and sustainability of new content on the platform. Indeed, while those projects required the intervention of computer scientists and graphic designers in order to create high-quality resources, @-MUSE aims at empowering musicians into creating and sharing their lessons by themselves, on the base of a common frame of reference: the musical score.

To do so, we designed a MSB. It consists in a set of annotated performances (specimen, or instance) each related to a given musical work (species, or class). This base can be used to compare various performances from music experts or students, and also to dynamically build new music lessons from the available content. To do so, we define a Musical Sign (MS) [28], as an object including a content (a musical performance or demonstration), a form (a score representing the played piece) and a sense (the background experience of the performer, what he or she intends to show) from the viewpoint of a subject (the creator of the Sign). Figure 5 describes the composition of a MS that can be shared on the platform through a multimedia annotation. Indeed, the principle of @-MUSE is to illustrate abstract scores with indexed multimedia content on top of MusicXML format [33] in order to explicit concretely how to interpret them. Besides, as shown on Figure 5, multimedia annotations embed all three components of a Sign (data, information and sense). This procedure is inspired from a common practice in the music education field, which consists in adding annotations on sheet music in order to remember tips or advice that were validated during the instrumental practice [29]. Figure 6 presents an example of such practice on an advanced piece for the piano, where annotations indicate tips to overcome technical and expressive difficulties, and underline points to improve for the learner. Grounding the @-MUSE service on this practice insures a transparent and natural usage for musicians who already annotate their scores by hand, and additionally enables them to show what they mean using multimedia features. As such, @-MUSE empowers musicians into creating their own interactive scores, using for instance mobile tablets equipped with webcams (@-MUSE prototype [30]). Naturally, our platform usage on mobile devices is particularly relevant as music is rarely practiced in a classroom, in front of a computer, but rather in informal situations (in front of a music stand, at home or with friends). Moreover, recent tablets featuring advanced tactile and multimedia characteristics facilitate the navigation within the score and the creation of high quality content on the platform.

Collaborative aspects are also essential in music learning, where one progresses by confronting his performances to others'. In this frame, managing Signs rather than Knowledge is particularly relevant, as there is no "absolute truth" in artistic fields: each interpretation can lead to technical discussions between musicians, and their negotiations should be illustrated with live performances to be shown, then understood. This is why we introduced the notion of Musical Message Board (MMB) in [28]. MMBs support discussions between musicians through a Glosses' system, leading to the creation of a thread of MS indexed on some parts of the score (a note, a musical phrase, a measure, etc.). In addition to these indexed multimedia annotations, what distinguishes this MS thread from a discussion on the piece is that each of the created Signs is indexed in context and can then be reused in different situations, for instance, on another piece of music presenting similar features. To do so, the MSB should be able to grasp the basic sense of the created MS, in order to organize itself, and provide advanced Sign sharing functionalities to users.

Collecting MS on different pieces of music enables also the illustration of significant descriptive logics in order to organize the MSB. Descriptive logics of the semiotic Web are more meaningful than description logics of the semantic Web because they bring human interpretations (psychological annotations) on top of symbolic representations (formal notations). Indeed, in musical education, understanding the structure of a work is an important key to play it correctly. Musicology provides a guide for the musician to explore the piece in the finest details and to better assimilate it. But this structure can be lively exemplified with MS created by @-MUSE in the signification process of understanding the context of resolution of the musical piece.

Descriptive logics express the background knowledge of specialists who well understand the historical context of music playing. It often depends on the style or form of the considered piece, i.e. its classification. For instance, a fugue is based on a theme that is repeated all along the piece in different voices [30]. Underlining these themes within the score allows disposing of a framework to better analyze the corresponding performances and establish fruitful confrontations. Figure 7 gives an example of an ontology based on descriptive logics (generic musical analysis).

This decomposition corresponds to the traditional way music teachers introduce a new piece to students [31]. After a short overview of the piece context (composer, style, mood), its characteristics patterns and difficult parts are identified and commented. This process can be recreated within the @-MUSE platform thanks to the characterization of descriptive logics adapted to each musical style. From the human-machine interaction point of view, it consists in proposing an "annotation guide" for each new piece, in order to obtain a complete interactive score at the end of the process. This method intends to guide users into the semiosis process described in Figure 2, by providing them a framework to communicate their own view of the considered piece and its characteristic features. In order to model these descriptive logics more formally, we proposed in [31] a Musical Performance Ontology based on the Music Ontology [32]. This ontology enables the automatic manipulation of concepts related to the piece structure, but also to gestural and expressive work. Tagging MS with these concepts and relations allows @-MUSE to automatically generate appropriate annotations on new pieces. Indeed, while machines can hardly deal with expressive and emotional information, they can provide basic information on specific patterns or unknown symbols, given a style or composer context. To do so, we designed a Score Analyzer [34] to automatically extract difficult parts within a given score, and to generate basic annotations. This prototype is based on the extraction of characteristic features of a score: chords, hands displacements, fingering, tempo, harmonies, rhythms and length. Work is in progress to measure the relevance of these estimations in comparison to human appreciations. As such, @-MUSE proposes an innovative service to share Musical Signs on a collaborative Web platform. The usage of multimedia and validated standards such as MusicXML empowers users into illustrating specific parts of a musical work in a collaborative and reusable way. Perspectives of this research include further testing with musician collaborators from music schools, as well as work on decision support for automatic score annotation.

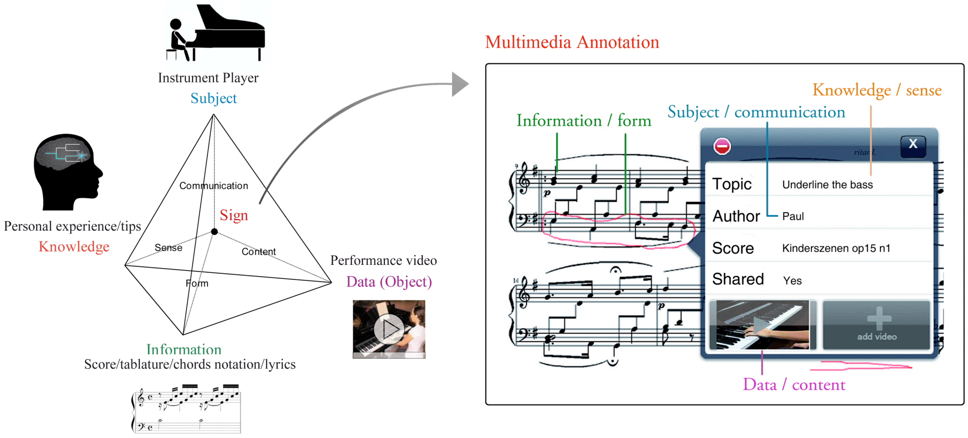


Figure 5. The musical sign tetrahedron illustrated with a multimedia annotation on @-MUSE

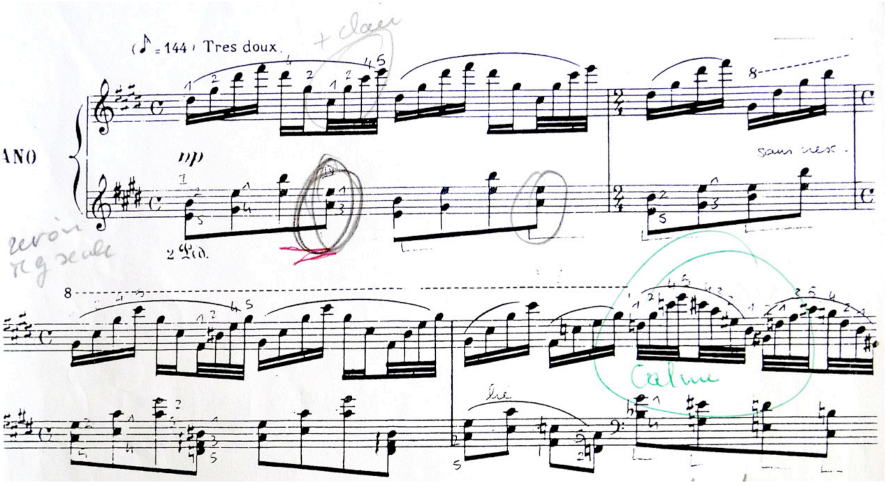


Figure 6. Annotated score example (extract from “Jeux d’Eau” by Maurice Ravel

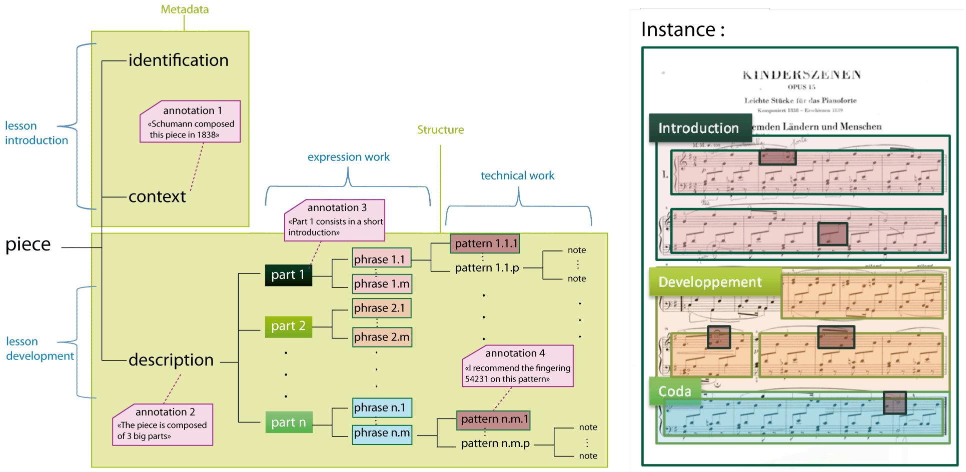


Figure 7. The musical descriptive model supporting descriptive logics of the Semiotic Web

6 SIGN SHARING IN BIODIVERSITY INFORMATICS

Sharing Signs is also relevant in biodiversity management with ICT, in this specific domain called Biodiversity Informatics that applies computerized acquisition and processing methods to natural data, information and knowledge, in order to define, describe, classify and identify biological objects. More precisely, we focus on the scientific discipline called *Systematics* that deals with listing, describing, naming, classifying and identifying living organisms. Our natural objects are living specimens in the fields and in museum collections. Experts in Systematics at university or in museums have studied them intimately for years and are able to recognize their names that give access to more information in monographs.

In this frame, the IKBS project (Iterative Knowledge Base System) aims at constituting a Sign Base (SB) rather than a Knowledge Base (KB) with the interactions coming from a community of biologists and amateurs that want to share their interpretations of observations. This project will benefit from the long experience we accumulated in the field of Mascarene Corals [35] and Plants identification [36]. Figure 8 sums up our Knowledge transmission process in this domain, based on a Creativity Platform.

It applies the experimental and inductive approach in biology, conjecture and test [37], with a natural process of knowledge management that is well suited to teaching from real examples:

1. Acquire	2. Process	3. Validate
● Define a descriptive model ● Build a multimedia questionnaire ● Describe classified cases (already identified)	● Classify cases by classification methods ● Discriminate them by decision trees ● Identify them with case-based reasoning (nearest neighbors)	● Refine cases and initial descriptive model ● Update the old cases with new observations ● Discover new knowledge (new classifications)

4. Iterate

Figure 8. Knowledge management cycle with IKBS

Indeed, IKBS has developped an original approach based on collection specimens' descriptions for helping specialists to discover new knowledge and classifications:

1. Acquisition of a descriptive model and descriptions,
2. Processing of this knowledge and case base for classification and identification purposes,
3. Experimentation, validation and refinement of cases and descriptive models, and
4. Iteration.

For identification purpose, the expert controls the transmission process, which is detailed in figure 9 for corals:

- The actual node of the decision tree or identification key is shown (e.g. inter-corallite's line),
- The referred question for this descriptive component and the illustrations of its possible values at this node are directly accessible (e.g. present or absent),
- One can leaf through the list of indexed cases at this node (e.g. cases of *Pocillopora* and *Stylophora*), in order to see the different values of the components and specimens,
- The pictures for the remaining objects at the current node are shown. The identification key may be useful to learn species' characteristics and improve one's ability to observe specimens in their natural surroundings or in a museum collection.

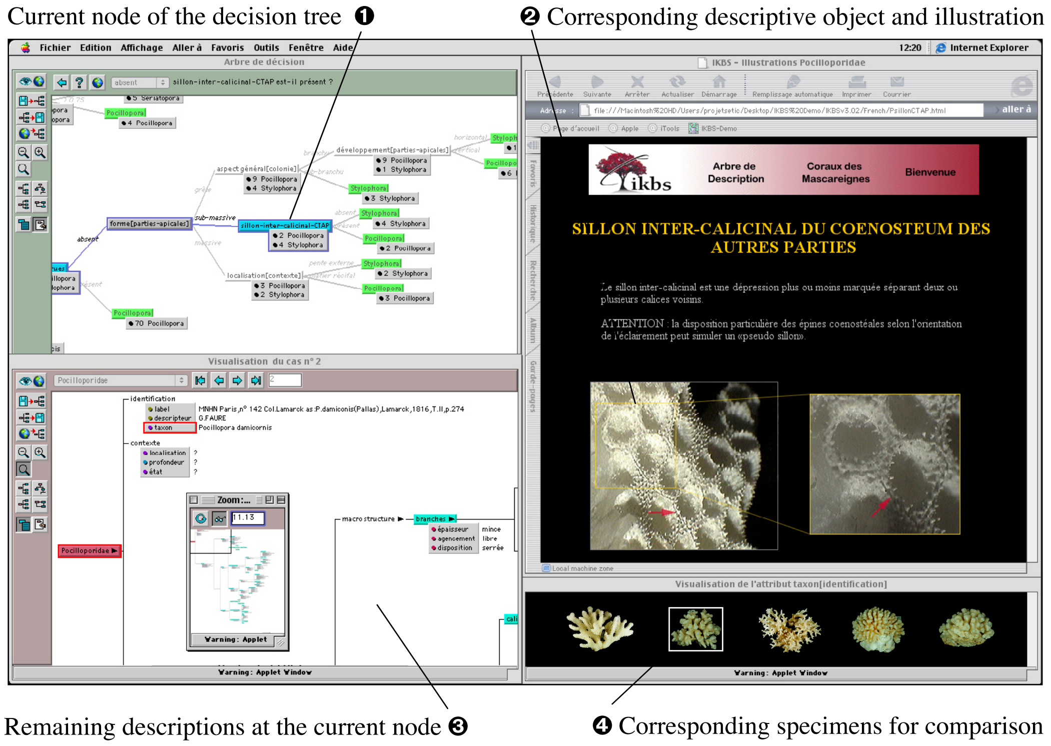
But the Learning problem from the end-user viewpoint is to *know how to observe* these objects in order to identify correctly the name of the species. This task is complex and needs help from the specialists who know by experience where to observe correctly the "right characters". By taking care of this knowledge transmission bottleneck, we enter the domain of Sign management for getting more robust results with end-users.

Our idea of Sign management is to involve end-users with researchers and entrepreneurs for making them participate to the design of the product/service that they want.

The problem that we have to face with when making knowledge bases is that their usefulness depends on the right interpretation of questions that are proposed by the system to obtain a good result.

Hence, in order to get correct identifications, it is necessary to acquire qualitative descriptions. But these descriptions rely themselves on the observation guide that is proposed by the descriptive model. Moreover, the definition of this ontology is dependent upon easy visualization of descriptive logics. At last, the objects that are part of the descriptive model must be explained in a thesaurus for them to be correctly interpreted by targeted end-users. Behind each Object, there is a Subject that models this Object and gives it an interpretation. In life sciences, these objects can be shown to other interpreters and this communication between Subjects is compulsory for sharing interpretations, and not only transmitting knowledge.

The challenge of Sign management for Science observation such as Systematics is to involve all types of end-users in the co-design of Sign bases for them to be really used (e-service). It is why we, as biologists and computer scientist (biomaticians), emphasize the instantiation of a Living Lab in Teaching and Learning at University of Reunion Island for sharing interpretations of objects and specimens on the table rather than concepts and taxa in the head of subjects (figure 10) : draw me a sheep, said the little prince !


Current node of the decision tree ❶
❷ Corresponding descriptive object and illustration
Fichier Edition Affichage Aller à Favoris Outils Fenêtre Aide
Arbre de décision
absent
forme[parties-apicales]
sub-massive
sillon-inter-calicinal-CTAP
aspect général[colonie]
développement[parties-apicales]
localisation[contexte]
Pocillopora
Stylophora
Visualisation du cas n° 2
Pocilloporidae
Zoom:...
macro structure
branches
Warning: Applet Window
12:20
Internet Explorer
IKBS - Illustrations Pocilloporidae
Précédente Suivante Arrêter Actualiser Démarrage Remplissage automatique Imprimer Courrier
aller à
ikbs
Arbre de Description
Coraux des Mascareignes
Bienvenue
Visualisation de l'attribut taxon[identification]
Remaining descriptions at the current node ❸
❹ Corresponding specimens for comparison

sillon-inter-calicinal-CTAP est-il présent ?

identification
label MNHN Paris,n° 142 Col.Lamarck as:P.damiconis(Pallas),Lamarck,1816,T.II,p.274
descripteur G.FAURE
taxon Pocillopora damicornis

contexte
localisation ?
profondeur ?
état ?

épaisseur mince
agencement libre
disposition serrée

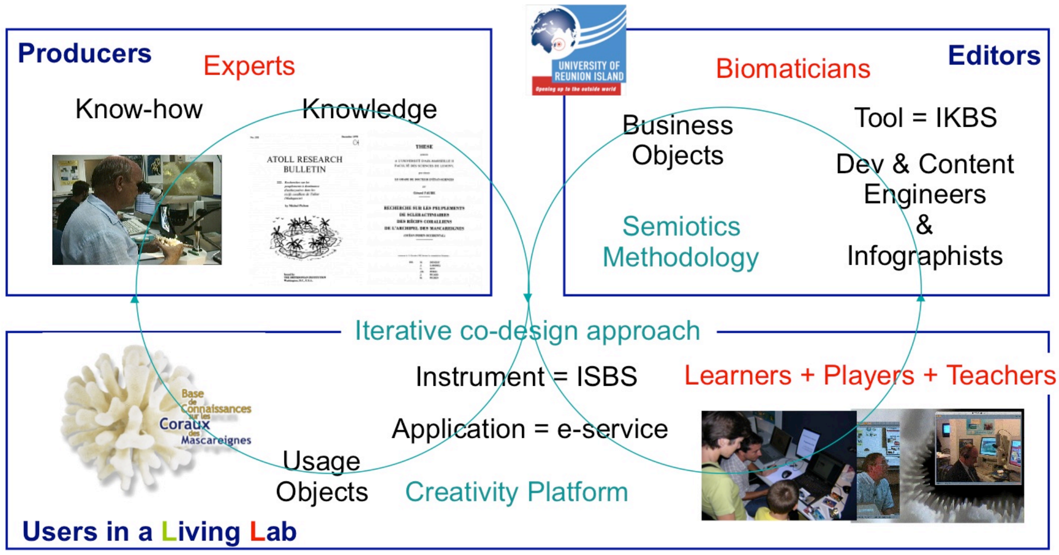
Adresse : file:///Macintosh%20HD/Users/projetsetic/Desktop/IKBS%20Demo/IKBSv3.02/French/PsillonCTAP.html

SILLON INTER-CALICINAL DU COENOSTEUM DES AUTRES PARTIES

Le sillon inter-calicinal est une dépression plus ou moins marquée séparant deux ou plusieurs calices voisins.

ATTENTION : la disposition particulière des épines coenostéales selon l'orientation de l'éclairement peut simuler un «pseudo sillon».

Figure 9. The identification process for corals


Producers
Experts
Know-how
Knowledge
Biomaticians
Editors
Business Objects
Tool = IKBS
Dev & Content Engineers & Infographists
Semiotics Methodology
Iterative co-design approach
Instrument = ISBS
Learners + Players + Teachers
Application = e-service
Usage Objects
Creativity Platform
Users in a Living Lab

UNIVERSITY OF REUNION ISLAND
ATOLL RESEARCH BULLETIN
THESE
Base de connaissances sur les Coraux des Mascareignes

Figure 10. The Sign management process for coral objects' interpretation

7 DISCUSSION

As shown in music and biodiversity Teaching and Learning, if we want to innovate with people, we should use the concept of Sign management rather than Knowledge management, because the paradigm shift is to pass from knowledge transmission to sign sharing by managing know-how.

Since several years in computer-aided systematics, we proposed a knowledge management methodology based on a top-down transmission of experts' knowledge, i.e. acquisition of a descriptive model and structured cases and then processing of these specimens' descriptions with decision trees and case-based reasoning. We designed a tool called IKBS for Iterative Knowledge Base System to build knowledge bases. But the fact is that Knowledge is transmitted with text, not shared with multimedia, and there is a gap between interpretations of specialists and end-users that prevents these lasts from getting the right identification.

More recently in instrumental e-Learning, we focused on the need to show gestural know-how with interactive multimedia contents to play correctly a piece of music, by annotating electronic scores with @-MUSE. This pedagogical approach is based on a gloss system on the Web that can be indexed in codified musical notation.

Today, we prefer to deliver a Sign management method for Teaching and Learning how to identify these collection pieces (specimens or scores) on a Co-Design or Creativity Platform. This bottom-up approach is more pragmatic and user-centered than the previous one because it implicates end-users at will and is open to questions and answers. The role of biological and musical experts is to show amateurs how to play, observe, interpret and describe these art and science works. The responsibility of semioticians (the new cognitians) is to store and share experts' interpretations of their observation and playing, i.e. know-how rather than knowledge in sign bases with multimedia annotations for helping them to define terms, model their domain, and allow end-users to interpret correctly the objects.

As computer scientists and knowledge engineers, we want to design a new Iterative Sign Base System (ISBS) that will be the kernel of our Information Service for defining ontologies and terms, describing pieces work, classifying them with machine learning techniques, and identifying the name through a multimedia interactive questionnaire. The objective of such a tool is to become an instrument in users' hands for monitoring biodiversity in the fields with the National Park of Reunion Island, and music at home with the Regional Music Conservatory.

For achieving this, we stressed on the importance of reducing the gap between interpretations of teachers (specialists) and learners (amateurs) to get the right identification name and then access to information in databases, or to get the correct gesture that gives the right sound for playing music. This pedagogical effort must concretize itself on a Co-Design or Creativity Platform, which is the Living Lab meeting place for teachers, players and learners, and where these people can manipulate the objects under study, test the proposed e-services and be guided by experts' advices. The teacher is a *producer* who communicates his skilled interpretation of an activity at different levels of perception: psychological motivation, training action, and reasoning feedback. The players are designers-developers *editors* that produce multimedia contents of the expert tasks to perform a good result and index them in a sign base. The learners are *prosumers* (producers and consumers) who experiment the sign bases on the physical or virtual Co-Design Platform and tell about their use of the tool to domain experts, ergonomists and anthropologists, in order to improve the content and the functionalities of the mock-ups and prototypes.

Behind each Object to observe, play and describe, there is a Subject who expresses himself and interprets an object by adding his proper signification. This is why we differentiate the Semantic Web, which is the business object approach (the Web of things) represented "objectively" with some description logics (formal syntax for ontologies and cases), and the Semiotic Web that is the usage object approach (the Web of Signs) signified by some descriptive logics of the domain (meaningful process of performance), and which are more subjective. The purpose of the Semiotic Web is to facilitate a consensus between community members, without forgetting that some interpreters are smarter than others in performing a Science or an Art. Their expertise will be visible if users show their interpretations of objects by multimedia artifacts (HD video, 3D simulation, annotated drawings or photos), and if other end-users can ask questions on their know-how and negotiate interpretations. It is why in the frame of natural and cultural heritage enhancement, we proposed to develop Teaching and Learning by Playing e-services with people in a Living Lab by using Sign management on a Co-design Platform at the University of Reunion Island [38].

8 CONCLUSION

In the post-industrial age of our digital society, designing new services on the Web is crucial for regional territories in order that they become more attractive, competitive, and also more sustainable in the global economy. But up to now, innovation is mainly seen as a linear technological downstream process, centered on enterprises (clusters) and not viewed as an iterative usage upstream process, focused on individuals (Living Labs).

The *form* of LL is attractive because it is an ecosystem based on democratizing innovation with people. User-centered design innovation means that some people, called lead-users, want to innovate for themselves. It has been shown that these persons make most of the design of new services, and only a few come from manufactures.

The *content* of LL is competitive because the best solutions from lead-users are experimented in real time by making situational analyses in "usage laboratories". Mock-ups and prototypes are tested and instrumented to get the best-customized-personalized products and services. For example, the game design (user interaction) and interfaces of 3D multimedia video games benefit greatly from the analysis of feedbacks coming from end-users in communities of practice. So, the success of the e-service does not depend only on the technical success: it has more to do with the quality of human-computer interaction provided with the technology.

At last, the *sense* of LL should be more sustainable, i.e. to render a useful and free service before being profitable, i.e. not only based on a monetary basis but also on trust and reputation. This characteristic is fundamental in the meaning of open access innovation to serve a mission within the scope of products and services made by publicly funded universities. The ultimate value would be to create a form of digital companioning in order to reposition human sharing at the core of technology race.

ACKNOWLEDGEMENTS

We would like to thank the French National Government, the Regional Council of Reunion Island and the European Union for supporting this research in the frame of the PO-FEDER 2-06 ICT measure: http://www.reunioneurope.org/UE_PO-FEDER.asp

REFERENCES

[1] http://en.wikipedia.org/wiki/Internet

[2] http://www.openlivinglabs.eu/news/when-infrastructure-meets-user-new-lovestory-making

[3] www-sop.inria.fr/teams/axis/LLSS2010/ecoleLL/content/literature

[4] Chesbrough, H. W.: Open Innovation: The New Imperative for Creating and Profiting from Technology, Boston, Harvard Business School Press, (2003).

[5] Von Hippel, Eric A.: Democratizing Innovation, MIT Press, Cambridge, MA (2005).

[6] Keen, A.: The Cult of the amateur: How Today's Internet is Killing Our Culture, Doubleday, New York, 2007, 228 pp, (2007).

[7] Lassila, O., Hendler, J.: Embracing "Web 3.0", IEEE Internet Computing, vol. 11, no. 3, pp. 90-93, doi:10.1109/MIC.2007.52, (2007).

[8] Peirce, C.S.: Elements of Logic, In Collected Papers of C. S. Peirce (1839 - 1914), C. H. Hartshone & P. Weiss, Eds. The Belknap Press, Harvard Univ. Press, Cambridge, MA, (1965).

[9] Uexküll, J. von.: Theoretical Biology. (Transl. by D. L. MacKinnon. International Library of Psychology, Philosophy and Scientific Method.) London: Kegan Paul, Trench, Trubner & Co. xvi+362, (1926).

[10] Farreny, H.: Les Systemes Experts: principles et exemples, Toulouse: Cepadues-Editions, (1985).

[11] Shapiro, J. A.: Bacteria are small but not stupid: cognition, natural genetic engineering and sociobacteriology, Studies in History and Philosophy of Biological and Biomedical Sciences, Vol 38, Issue 4, p. 807–819, (2007).

[12] Maslow, A.H.: A Theory of Human Motivation, Psychological Review Vol 50, No 4, p. 370-96, (1943).

[13] Engeström, Y.: Learning by expanding: an activity-theoretical approach to developmental research, Orienta-Konsultit Oy, Helsinki, (1987).

[14] Barbieri, M.: Introduction to Biosemiotics. The new biological synthesis. Springer, (2007).

[15] Sebeok, T.A.; Umiker-Sebeok J. (eds.) Biosemiotics: The Semiotic Web 1991. Berlin: Mouton de Gruyter, (1992).

[16] Saussure de, F.: Nature of the Linguistics Sign, in: Charles Bally & Albert Sechehaye (Ed.), Cours de linguistique générale, McGraw Hill Education, (1916).

[17] University of Reunion Island Living Lab vision (2011): http://www.slideshare.net/conruyt/urlltl

[18] New Media Consortium (2011): http://wp.nmc.org/horizon2011/

[19] Conruyt, N., Sebastien, D., Vignes-Lebbe, R., Cosadia, S., Touraivane,: Moving from Biodiversity Information Systems to Biodiversity Information Services, Information and Communication Technologies for Biodiversity and Agriculture, Ed by : L. Maurer and K. Tochtermann, ISBN: 978-3-8322-8459-6, Shaker Verlag, Aachen, (2010).

[20] Vibrant FP7 Project (2011): http://vbrant.eu/

[21] Conruyt, N., Grosser, D.: Knowledge management in environmental sciences with IKBS: application to Systematics of Corals of the Mascarene Archipelago, Selected Contributions in Data Analysis and Classification, Series: Studies in Classification, Data Analysis, and Knowledge Organization, pp. 333-344, Springer, ISBN: 978-3-540-73558-8, (2007).

[22] Le Renard, J., Conruyt, N.: On the representation of observational data used for classification and identification of natural objects, LNAI IFCS'93, 308–315, (1994).

[23] Conruyt, N., Sébastien, O., Sébastien, V. Sébastien, D. Grosser, D., Calderoni, S., Hoarau, D., Sida, P.: From Knowledge to Sign Management on a Creativity Platform, Application to Instrumental E-learning, 4th IEEE International Conference on Digital Ecosystems and Technologies, DEST, April 13-16, Dubaï, UAE, (2010).

[24] Ryle, G.: The concept of mind. London: Hutchinson, (1949).

[25] Callaos, N.: The Essence of Engineering and Meta-Engineering: A Work in Progress, The 3rd International Multi-Conference on Engineering and Technological Innovation: IMETI 2010, June 29-July 2, Orlando, Florida, USA, (2010).

[26] www.slideshare.net/conruyt/living-lab-and-digital-cultural-heritage

[27] Sébastien, O., Conruyt, N., Grosser, D.: Defining e-services using a co-design platform: Example in the domain of instrumental e-learning, Journal of Interactive Technology and Smart Education, Vol. 5, issue 3, pp. 144-156, ISSN 1741-5659, Emerald Group Publishing Limited, (2008).

[28] Sébastien, V., Sébastien, D., Conruyt, N.: Dynamic Music Lessons on a Collaborative Score Annotation Platform, The Sixth International Conference on Internet and Web Applications and Services , ICIW 2011, St. Maarten, Netherlands Antilles, pp. 178-183, (2011).

[29] Winget, M. A.: Annotations on musical scores by performing musicians: Collaborative models, interactive methods, and music digital library tool development, Journal of the American Society for Information Science and Technology, (2008).

[30] Sébastien, V., Sébastien, P., Conruyt, N.: @-MUSE: Sharing musical know-how through mobile devices interfaces, 5th Conference on e-Learning Excellence in the Middle East, Dubaï, (2012).

[31] Sébastien, V., Sébastien, D., Conruyt, N.: An Ontology for Musical Performances Analysis. Application to a Collaborative Platform dedicated to Instrumental Practice, The Fifth International Conference on Internet and Web Applications and Services, ICIW, Barcelona, pp. 538-543, (2010).

[32] Raimond, Y., Abdallah, S., Sandler, M., Giasson, F.: The Music Ontology, Proceedings of the International Conference on Music Information Retrieval, ISMIR, (2007).

[33] Castan, G., Good, M., Roland, P.: Extensible Markup Language (XML) for Music Applications: An Introduction, The Virtual Score: Representation, Retrieval, Restoration, MIT Press, Cambridge, MA, pp. 95-102, (2001).

[34] Sébastien, V., Sébastien, D., Conruyt, N.: Constituting a Musical Sign Base through Score Analysis and Annotation, The International Journal On Advances in Networks and Services, (2012).

[35] http://coraux.univ-reunion.fr/

[36] http://mahots.univ-reunion.fr

[37] Popper, K.R., La logique de la découverte scientifique, Payot (Eds.) Press, Paris, (1973).

[38] http://www.openlivinglabs.eu/livinglab/university-reunion-island-living-lab-teaching-and-learning

Ontology Learning From Unstructured Data for Knowledge Management: A Literature Review

Jens Obermann and **Andreas Scheuermann**[1]

Abstract. In the global race for competitive advantage Knowledge Management gains increasing importance for companies. Purposefully creating and exploiting knowledge demands advanced Information Technology. Since ontologies already constitute a basic ingredient of Knowledge Management, approaches for ontology learning form unstructured data have the potential to bring additional advantages for Knowledge Management. This work presents a study of state-of-the-art research of ontology learning approaches from unstructured data for Knowledge Management. Nine approaches for ontology learning from unstructured data are identified from a systematic review of literature. A six point comparison framework is developed. The comparison results are analyzed, synthesized, and discussed to inform future research on approaches for ontology learning from unstructured data for Knowledge Management.

1 Introduction

In the global race for competitive advantage the ultimate success of companies increasingly depends on effectively and efficiently exploiting knowledge. Knowledge Management (KM) aims at creating and exploiting knowledge in a purposeful, structured, and organized manner. Thereby, KM depends on non-technical aspects but increasingly on advanced Information Technology (IT). IT supports not only single aspects of KM but rather KM as a whole. From an IT perspective, ontologies play a significant role to support Knowledge Management [1][2].

However, building ontologies (from scratch) is a non-trivial, complex, cumbersome, time-consuming, labor-intensive, and error-prone task. It typically involves experts from the area of ontology engineering and experts from the particular domain of interest. However, such experts are scarce and it is hard to elicit the relevant knowledge of a human expert. In total, ontologies constitute an essential component of Knowledge Management but building, extending, and refining ontologies induces lots of efforts and costs for companies.

To overcome these obstacles, ontology learning represents an approach to (semi-)automatically build, extend, and refine ontologies. Ontology learning bears the potential to leverage ontologies for KM without inducing excessive costs. As such, ontology learning appears to be somewhat like the ideal solution for ontologies for Knowledge Management. Unfortunately, huge amounts of knowledge relevant for companies lack a clear structure and organization [3]. This unstructured knowledge aggravates (semi-)automatically processing by machines and, thus, drastically increases the number of errors as well as demands for human intervention. For instance, particularly, the strong adoption of the idea of Social Media on the World Wide Web (WWW) increased the number of unstructured data sources dramatically. Social Media in terms of Facebook, Twitter, Blogs, and further types of these applications contain plenty of unstructured data, which can be of major significance for companies, e.g., marketing, product development, consumer studies, customer relationships, advertising, recruiting, etc. As a consequence, it is and becomes more and more essential for companies to exploit unstructured data sources in order to improve their position in the competitive market environment.

This paper reviews and analyzes existing approaches for ontology learning form unstructured data for Knowledge Management. The goal is to inform future research in the area of ontology learning from unstructured data for Knowledge Management. This work extends and refines [4][5] particularly with respect to unstructured data and Knowledge Management.

The structure of the paper is as follows: Section 2 introduces basic concepts from the area of ontologies, and ontology learning. Section 3 presents the ontology learning approaches for unstructured data resulting from literature analysis. Section 4 characterizes the comparison framework for analysis whereas Section 5 presents the analysis results and discusses them. Section 6 draws a conclusion.

2 Ontology and Ontology Learning for Knowledge Management

2.1 Ontology

Artificial Intelligence (AI) provides several definitions of the term ontology. The most prominent one stems from [6][7]. Accordingly, “ontology is an explicit specification of a conceptualization”. [8] extend this definition, requiring the specification to be formal and the conceptualization to be shared. Moreover, [9] defines an ontology as “a formal, explicit specification of a shared conceptualization of a domain of interest” where

- conceptualization conforms to an abstract model of some (real-world) phenomenon by having determined its relevant concepts,
- explicit depicts that the type of concepts and the constraints holding on their use are explicitly defined.
- formal refers to the fact that the ontology should be machine-readable (automatically processed by machines) and, thus,

[1] Information Systems 2, University of Hohenheim, Stuttgart, Germany, email: jens.obermann@uni-hohenheim.de; andreas.scheuermann@uni-hohenheim.de;

interpretable by machines (e.g., which excludes natural language).

- shared reflects the notion that an ontology captures consensual knowledge being not private to an individual person, but accepted by a group of individuals.

To conclude, this work adopts the definition proposed by [9], because it is both comprehensive and concise.

Ontologies for Knowledge Management bear the potential to provide support for [7][9][10][11][12]:

- formally specifying knowledge about a specific domain of interest,
- structuring and organizing this knowledge,
- establishing a common terminology, i.e. interlingua,
- semantic interoperability and data integration, and
- sharing and reusing knowledge.

Against this background, ontologies provide dedicated means for the purpose of Knowledge Management.

2.2 Ontology Learning

2.2.1 Characteristics

Ontology learning is a relatively new and interesting research field in the area of AI and, particularly, in the intersection of information extraction and ontology engineering.

Ontology learning aims at (semi-)automatically acquiring knowledge from knowledge sources to create, i.e. build, extend, and refine ontologies [13]. Thereto, ontology learning mainly relies on extracting information, patterns, and relations from various kinds of knowledge sources, e.g., unstructured data [14]. The main advantages of (semi-)automatically engineering ontologies are reduced costs, time efforts, and errors due to less human interaction. In addition, limiting human interaction in the ontology engineering process may result in more suitable ontologies with respect to specific applications, e.g., Knowledge Management [14].

From a more technical point of view, ontology learning comprises two constituent components: (1) information extraction approaches and (2) learning approaches.

First, ontology learning distinguishes between two distinct approaches for information extraction: (1) rule-based approaches and (2) heuristics pattern approaches [13]. Both approaches build on lexico-syntactic patterns. These patterns allow for dealing with situations, i.e. knowledge sources characterized by insufficient pre-encoded knowledge and wide ranges of data (e.g., text) [15]. Lexico-syntactic patterns aim at finding recognizable structures within data (e.g., text) without depending on sets of fixed terms of expressions, which have to be determined in advance.

Second, ontology learning relies on learning algorithms. Learning algorithms decide whether extracted information entities fit the respective ontology. In case of a positive assessment, the algorithm adds a new element in terms of classes, properties, or axioms to the ontology [5][16].

Basically, ontology learning differentiates from other existing approaches by its multidisciplinary applicability and the potential to exploit vast and heterogeneous sources of data [17]. As such, ontology learning is both applicable to Knowledge Management and could improve it significantly.

2.2.2 Classification

Literature proposes several criteria to classify ontology learning approaches. Ontology learning approaches deal with an extraction of information from various types of knowledge sources ranging from structured data to unstructured data [5]. Accordingly, [18] use the type of knowledge sources to classify ontology learning approaches. The classification builds on free texts, dictionaries, semi-structured schemata, and relational schemata. Each of the knowledge sources requires distinct ontology learning approaches and techniques.

In essence, this work adopts the classification proposed by [18] but in a slightly different variant. The classification of ontology learning approaches of this work basically relies on the following types of knowledge sources:

- *structured data* are tightly coupled to specific rules of a conceptualization, e.g., relational databases and the relational schema.
- *semi-structured data* incorporate some rules of conceptualization but also contain unstructured elements, e.g., HTML documents
- *unstructured data* can be of any kind and do not follow any rules or structure. The main characteristic of unstructured data is the high availability throughout all domains but also the lowest accessibility for ontology learning.

Considering these three types of knowledge sources, this work focuses on semi-automated and automated ontology learning approaches from unstructured data.

3 Ontology Learning from Unstructured Data

Ontology learning from unstructured data distinguishes between two different approaches: (1) statistical and (2) natural language processing (NLP). The literature review reveals four statistical ontology learning approaches and five ontology learning approaches based on NLP. An overview of both these ontology learning approaches is provided the following two subsections. It serves as an introduction to their analysis in Section 5.

3.1 Statistical Approaches

Since statistical ontology learning approaches in this subsection deal with unstructured data, they build on a common basic assumption. This basic assumption corresponds to the distributional hypothesis. The distributional hypothesis states that similar words often occur in similar contexts and it utilizes statistical patterns, which give hints for certain relations between words [19].

The automatic ontology learning approach proposed by [20] deals with enriching the WordNet database by using unstructured data sources of the WWW. An enrichment of WordNet is supposed necessary because of shortcomings with respect to (1) semantic variant concepts of words, which are related by topics, are not interlinked (e.g., to paint and paint or sun cream and beach) and (2) the vast collection of word meanings without any clear distinction. In this context, the proposed ontology learning approach uses word lists. Word lists describe the sense of the words of interest. This is based on the idea that specific other words describe the context and, thus, express the meaning of the word of interest. Initially, the proposed approach queries (boolean search query) the WWW for documents (datasets) containing the word of interest. The higher the number of the words of interest in the retrieved documents, the higher the statistical likelihood that the document correlates with the searched topic. To increase this likelihood, other descriptive words can be explicitly excluded from the search. Then, counting the appearance of single words in the documents and applying calculated distance metrics to hierarchically sort them results in topic signatures. These topic signatures are clustered and evaluated by means of a disambiguation algorithm. Thereby, clustering corresponds to a common technique to generate prototype-based, hierarchical ontologies. A defined semantic distance algorithm works as a measurement to agglomerate terms or clusters of terms. The largest or the least homogeneous cluster is split into smaller sub-groups by a divisive process to refine the ontology.

[21] propose a semi-automatic approach for enriching existing ontologies. The proposed approach is illustrated with an example of a medical ontology. Similar to [20], [21] utilize a statistical approach to cluster words occurring in a certain context to each other and sets of predefined rules. These rules represent distance measurements, e.g., maximum word distance between two words in a document. Thereby, the rules should not be contradictory to already existing distance measures. The statistical similarity measurement relies on the Kullback-Leibler divergence [22], which was originally designed to test the probability distribution between statistical populations in terms of information. [21] use the Kullback-Leibler divergence to check the weighted probability of a given linguistic property *w* with respect to the fulfillment by a word *x*. This allows for assessing and minimizing the distance of the word of interest and the retrieved document in a way similar to an optimization problem.

[23] also exploit web documents as a source to automatically create ontologies. This is because using the WWW for ontology creation might increase the probability that the ontology is up to date and more complete. Similar to [20] and [21], [23] create queries to search for specific words of interest and constraint search by criteria like the maximum number of returned results as well as using a filter for similar documents. Based on the initial results, a first analysis according to defined prerequisites is conducted in order to filter relevant documents. Then, utilizing statistical analysis aims at further filtering the most relevant documents from this subset. The next step is to filter the results from the previous step by using a new search word refining the original one. The last two steps mainly increase search depth. The resulting taxonomies support finding new relations between ontologies.

[24] adopt Formal Concept Analysis (FCA) for ontology learning on a completely automated basis. The proposed ontology learning approach analysis documents by searching for sets of object attributes and, based on them, derives relationships and dependencies between the objects. The results conform to nouns associated with several verbs as trailed attributes. These attributes define the context of the noun. The formal abstraction of the inherited nouns provides additional benefits to an end-user as the verbs enrich the created ontology. The reason for this may be the more adequate description provided by a verb in contrast to a noun hyponym.

3.2 NLP Approaches

Before introducing ontology learning approaches using natural language processing techniques, it is important to note that there exists not a clear and commonly agreed distinction between statistical and NLP approaches for ontology learning. Despite ontology learning approaches use statistics and linguistics to exploit unstructured data sources, NLP approaches incorporate a more intuitive way of dealing with unstructured data sources by using pattern recognition [25]. In particular, NLP approaches provide additional benefits with respect to knowledge-intensive domains making several constraints and rules necessary during ontology learning. Such constraints and rules conform to lexical inventories, syntactic rules, or previous defined conceptual knowledge [26].

[15] introduces the lexico-syntactic pattern extraction method to support the enrichment of existing patterns within the WordNet database by searching large text corpora as a mining resource for suitable semantic patterns. Crucial to this approach is that the English language has identifiable lexico-syntactic patterns, which indicate specific semantic relations as *is-a* relations. In comparison to other approaches, the text corpora have to fulfill only very little requirements for usage. In particular, this means that only one instance of a relation has to be available in the data source in order to decide whether the data source is suitable. [15] utilizes a deterministic system to provide one or several hypernyms for each unknown concept, which all have a certain probability to be correct from unstructured data. To increase the suitability of the derived concepts, the lexico-syntactic patterns have to fulfill some criteria. The lexico-syntactic patterns have to frequently occur in the text corpora, indicate the relation of interest, a possible recognition without any prior knowledge of the domain. Moreover, this approach allows for a combination with other techniques such as statistical algorithms for the purpose of refining the patterns found.

[27] build on the approach proposed by [15] in order to introduce an ontology learning approach to semi-automatically build an ontology from text corpora retrieved from a corporate intranet. The proposed approach incorporates a learning method, which is based on a set of given core concepts similar to WordNet. This learning method further uses statistical and pattern-based approaches to refine the result. The resulting ontology is pruned and restricted. In comparison to prior approaches, this (non-taxonomic) approach uses conceptual relations rather than manually encoded rules for the purpose of ontology creation. Moreover, this approach comprises a set of evaluation metrics to ensure a hegemonic ontology with respect to the target structure. However, these metrics are not conclusive enough to fully automate the ontology learning process.

[28] propose an approach to speed up the ontology learning process by using WordNet and, particularly, by creating sublanguages, i.e., WordNets. A creation of these sublanguages results from an application of Acronym Extractors and Phrase Generators in order to analyze data sources for concept elements. For instance, concept elements are words, potential relations between words, and phrases. Potential relationships are analyzed again and proposed for being added to the ontology. Words and suitable relationships are clustered into groups and linked to the corresponding synsets in WordNet as SubWordNets. Finally, the last step focuses on maintenance of the retrieved concept elements and structures.

[29] introduce an automatic approach to extend a lexical semantic ontology. The proposed algorithm searches the existing ontology for similarity to a synset for information extraction. For this purpose, [29] define several signatures for the word of interest, which are evaluated with respect to their semantic similarity to existing words within the ontology. The signatures conform to: (1) topic signatures defining a list and frequency of co-occurring words, (2) subject signatures, which inherit a list of co-occurring verbs, (3) object signatures, which contain a list of verbs and prepositions, (4) and modifier signatures, which consist of adjectives and determiners Thereby, similar words should be assigned with similar signatures because they are represented in similar contexts. All the signatures are aggregated to an overall similarity value. For assessing the frequency, [29] use the method of [20] to achieve more accurate results. Thereby, the plain frequencies are changed into weights, which assess the support of a word in a specific context of a synset.

More recent research in the area of NLP extends the area from which text corpora are derived from. In this context, [30] introduce an approach to extract information directly from Twitter messages. This approach refines the attempt to retrieve information from unstructured data to a new and even more demanding level. Besides the challenge of retrieving the correct semantic relation of the sources, the problems of misspellings, abbreviations and colloquial speech in the Tweets occurs. Therefore, a normalization of the text is necessary before the process of extending or refining the ontology can take place. Besides enriching an ontology, annotations from the Twitter messages are included in the retrieved semantic structures. These annotations refer to as contextual relations like opinions or.

Table 1 presents the identified approaches for ontology learning from unstructured data (see Annex).

4 Comparison Framework

The comparison framework contains six criteria to analyze the identified ontology learning approaches from unstructured data. These six criteria stem from an analysis of literature of the area of ontology learning and are assessed relevant for reviewing ontology learning approaches with respect to Knowledge Management. In particular, the various criteria focus both on methodological aspects of the ontology learning approaches (Criteria 1-5) as well as on the resulting ontologies (Criterion 6).

Criterion 1 (Goal) aims to detect and analyze the primary goal of ontology learning in terms of the specific problem and the concrete context of application.

Criterion 2 (Methodology) aims to detect and analyze the methodological approach underpinning ontology learning from unstructured data. This criterion pays special attention to statistical and NLP approaches for ontology learning as they frequently occur in literature and can be assessed promising for Knowledge Management.

Criterion 3 (Technique) aims to detect and analyze the technique for extracting information. This criterion specializes Criterion 2. For instance, it elaborates on the process of information extraction or learning.

Criterion 4 (Degree of Automation) aims to detect the degree of automation the ontology learning is supposed to operate. The criterion degree of automation basically distinguishes between semi-automatic and (fully) automatic approaches for ontology learning. The degree of automation can be related to economical advantages, which gain importance in business contexts.

Criterion 5 (Ontology Reuse) aims to detect whether the ontology learning approach reuses existing (formal) bodies of knowledge, e.g., WordNet. This is especially interesting for KM as it can be assumed that there already exist ontologies, which have to be extended or refined. This criterion also expresses the basic understanding of the ontology learning approach: building up ontologies from scratch or finding an existing framework as a starting point.

Criterion 6 (Ontology Elements and Structure) aims to detect and analyze the ontology elements, which result from information extraction as subjects of learning. Ontology elements conform to nouns or characteristics such as relationships or taxonomies when applying more advance methods. Also, this criterion includes the conceptual structure acquired. This criterion stresses the importance of the achieved results of the ontology learning approach and is dependent on the results of Criterion 3 and as well of Criterion 5 whether structure and objects are pre-defined by reusing an existing ontology. This is of particular interest because it provides information what results Knowledge Management can expect and how results can be used for specific goals of Knowledge Management.

Having introduced the criteria constituting the comparison framework for assessing the ontology learning approaches, it is important to note that these criteria are extrinsic in their nature and allow for an assessment from an objective point of view.

Table 2 illustrates the comparison criteria for analyzing the identified ontology learning approaches with regard to Knowledge Management (see Annex).

5 Results and Discussion

This section synthesizes, summarizes, and discusses the major analysis results concerning ontology learning approaches from unstructured data for Knowledge Management.

5.1 Statistical Approaches

[20] focus on the problem of word ambiguity in order to enhance ontology learning. To enrich ontologies, [20] use Harris' distributional hypothesis to measure the relevance of the retrieved knowledge elements. These knowledge elements conform to concepts and relationships. The proposed methodology can be performed automatically and exploits WordNet. However, the results appear to be biased since the resulting ontology remains completely without supervision.

[21] use clustering techniques and measurements of similarity to retrieve nouns and relationships from retrieved knowledge sources. However, instead of using randomly assigned knowledge sources, [21] define several data sources, e.g., documents, which already deal with the topic and let the automatic algorithm operate on them. Consequently, the proposed approach can be recognized as a semi-automated approach, since suitable data sources are selected prior by hand. The proposed approach mainly reuses medical ontologies with respect to the reported applications.

[23] introduce an approach that is characterized by the creation of new ontologies and the enrichment existing ontologies by including additional knowledge, e.g. from the WWW. In general, the authors use key words for the first step. The algorithm focuses

mainly on deriving classes and is supposed to operate automatically.

[24] use FCA to create ontologies from scratch. This approach also is builds on Harris' distributional hypothesis. The proposed approach allows for deriving concepts and taxonomies from the data sources. However, the data sources have to be manually selected according to the target topic. As such, the approach operates semi-automatically.

5.2 NLP Approaches

The approach proposed by [15] aims at enriching existing ontologies by exploiting knowledge sources not limited to specific aspects. [15] uses lexico-syntactic patterns, nouns, and hyponym-hypernym relationships for retrieval and reuse WordNet. The approach is supposed to work in an automatic way.

[27] incorporate an application-driven perspective by targeting at companies' intranets as a data source for ontology learning. The approach is capable to retrieve *is-a* relationships and concepts from corporate intranets. Because of the nature of the intranets, this approach operates only semi-automatically. It reuses GermaNet and WordNet as ontologies.

[28] build on WordNet, but also create sublanguage WordNets to enrich an upper ontology. The rationale is to discover and identify pre-defined web documents to create WordNets. This allows for deriving complete WordNets assuming that the domain expert has selected suitable data source. This approach is capable of updating single WordNets without the need of dealing with the entire ontology. Nevertheless, this approach needs intervention and can only be performed in a semi-automatic way.

[29] aims at enriching existing ontologies of a domain of interest by exploiting pre-defined data source. The proposed approach works automatically without supervision as it generates information signatures based on a set of criteria. However, this approach requires large data sources to generate adequate signatures for a unique identification of the content elements.

[30] also have the goal to enrich existing ontologies by exploiting Twitter. The derived ontologies can be enriched with annotations and contextual relationships depending on accompanying words or hash-tags in the Twitter feeds. This approach is supposed to operate completely automatic.

5.3 Discussion

The two different approaches – statistical and NLP – primarily aim at enriching existing ontologies. This means that the respective approach demands for a basic structure to operate on and the approach qualitatively improves the underlying ontology. Exceptions of this are the approaches of [23][24], which both focus on statistical approaches. This circumstance may be due to higher likelihood of errors if ontologies are built from scratch since no clear guideline is provided and there exist no possible comparisons to similar structures. Nevertheless, approach of [28] appears to be ambiguous because [28] create discrete ontologies with the sublanguages WordNets but these synsets have to be connected to an upper ontology.

Multiple early approaches construct *is-a* relationships between nouns. This results rather in taxonomies than ontologies. Other types of relationships such as *part-of* relationships seem to be neglected. In contrast, the approaches proposed by [30] enrich the retrieved relationships with additional attributes derived from the data sources. Thereby, [30] additionally uses annotations by means of an extrapolation of contextual relationships. Considering the different methodologies with respect to a temporal dimension, it appears that the approaches evolved from simple relationship identification to more advanced concept recognition of the underlying data sources.

Not all authors publish excessive information about the performance of their methodologies but rather deliver the application of the proposed methodology as a proof of concept. Hence, an evaluation of the methodologies of [23], [15], [28], and [30] is not possible. Nevertheless, the remaining authors provide some more detailed information about their results. [20] assess their algorithm with respect to different levels of granularity. They provide information about how the algorithm performs on word disambiguation with the generated topic signatures. The results show that the algorithm performs best with respect to a coarse granularity. In contrast, with a keen granularity, the performance declines very steep. [21] report on similar results. Whereas the enrichment of some concepts performs well, i.e. propositions are added to concepts; other concepts are not enriched at all. This might be due to the greater potential of distributional meaning of some concepts, e.g., medical doctor is too general to achieve suitable results. [24] provide excessive statistical performance measurements of their clustering approach. Thereby, the results show that their approach achieves a slightly better (approx. 1%) performance with respect to retrieved concepts and precision than comparable approaches on two different domains. Another further advantage of this approach is that the concepts provide some additional description, which supports users to understand the retrieved ontology. [27] use a pattern-based approach to achieve a basic ontological structure and reached 76.29 % of correctly discovered relationships. Additionally, almost 50% of all dictionary entries are correctly imported into the ontology. [29] compare different methodologies of signature creation. The results show that only a combination of different signature methodologies for signature creation generates acceptable results in terms of accuracy.

It can be concluded that only some of the approaches provide (fully) automated and unsupervised ontology learning. The approaches generate basic concepts and provide relationships between derived concepts, but there is still manual intervention needed to complete the ontologies. Manual interventions occur in terms of a selection of the data source or the manual assessment of the retrieved relationships by a domain expert. This aggravates a usage on a larger scale because it is time and resource consuming. Only an automated and integrated step performing quality control allows the approaches to be used in an up-scaling context such as Knowledge Management.

Table 3 provides an overview of the results of the review of ontology learning approaches from unstructured data for Knowledge Management (see Annex).

6 Conclusion

To inform future research in the area of ontology learning from unstructured data for knowledge management, this work analyses and reviews existing ontology learning approaches either based on statistics or natural language processing. On the basis of a literature review, we identified and characterized four statistical approaches for ontology learning and five natural language processing approaches for ontology learning. To analyze these ontology learning approaches, six criteria originating from literature constitute the comparison framework. The results provide an overview of state-of-the-art research in the area of ontology learning from unstructured data, raise issues, and point to potential future aspects of research. The general conclusion is that approaches for ontology learning from unstructured data have a significant potential to improve Knowledge Management. However, we are still far away from realising this potential. This work is a rigorous and systematic attempt to identify, analyze, and synthesize the research in the area of approaches for ontology learning from unstructured data in the area of Knowledge Management.

Future work focuses on extending the comparison framework by adding further criteria dealing with aspects of both ontology learning and Knowledge Management. The goal is to elaborate in a more precise and fined-grained way on the potentials of ontology learning from unstructured data for Knowledge Management.

Acknowledgement

The work presented in this paper was partly funded by the German Federal Ministry of Education and Research under the project InterLogGrid (BMBF 01IG09010E).

REFERENCES

[1] S. Staab ´Wissensmanagement mit Ontologien und Metadaten'., *Informatik-Spektrum*, **25**(3), 194-202, (2002).

[2] S. Staab, H. Schnurr, R. Studer and Y. Sure, 'Knowledge Processes and Ontologies', *IEEE Intelligent Systems,* **16** (1), 26-34, (2001).

[3] D. McComb *Semantics in business systems*, Morgan Kaufman, Massachusetts, 1st edn., 2004.

[4] M. Hazman, S.R. El-Beltagy, S. Rafea, 'A Survey of Ontology Learning Approaches', *International Journal of Computer Applications,* **20** (9), 36-43, (2011).

[5] C. Biemann, 'Ontology Learning from Text: A Survey of Methods', *LDV Forum,* **20**(2), 75-93, (2005).

[6] T.R. Gruber, 'A Translation Approach to Portable Ontology Specifications', *Knowledge Acquisition,* **5**(2), 199-220, (1993).

[7] T.R. Gruber, 'Toward Principles for the Design of Ontologies Used for Knowledge Sharing', *International Journal of Human-Computer Studies,* **43**(5/6), 907-928, (1995).

[8] W. Borst, *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*, PhD Thesis, University of Enschede, 1997.

[9] R. Studer, R. Benjamins and D. Fensel, 'Knowledge Engineering: Principles and Methods', *Data and Knowledge Engineering,* **25**(1-2), 161-197, (1998).

[10] E. Motta, *Reusable Components for Knowledge Modeling*. Case Studies in Parametric Design. IOS Press, Amsterdam, 1999.

[11] N. Guarino, 'Formal Ontology and Information Systems', *Proceedings of the First International Conference on Formal Ontology in Information Systems,* Trento, Italy, (1998).

[12] M. Uschold 'Building ontologies: Towards a unified methodology', *Proceedings of Expert Systems '96, the 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems*, Cambridge, UK, (1996).

[13] R. Navigli, P. Velardi, S. Faralli, 'A Graph-Based Algorithm for Inducing Lexical Taxonomies from Scratch', *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, 1872-1878, (2011).

[14] M. Sabou et al., 'Learning domain ontologies for Web service descriptions: an experiment in bioinformatics'. *Proceedings of the 14th International Conference on World Wide Web*, 190-198, (2005).

[15] M.A. Hearst, 'Automatic Acquisition of Hyponyms from Large Text Corpora', *Proceedings of Coling*, 539–545, (1992).

[16] P. Cimiano et al., *Ontology Learning*, Handbook on Ontologies, Ed. by S. Staab, 2nd edn, Springer, Heidelberg, 2009.

[17] P. Buitelaar, P. Cimiano, B. Magnini, 'Ontology Learning from Text: An Overview', *Ontology Learning from Text: Methods, Evaluation, and Applications,* Ed. by B. Magnini, P. Buitelaar, P. Cimiano, Frontiers in Artificial Intelligence and Applications (123), 170-173, (2005).

[18] A. Mädche, S. Staab, 'Learning Ontologies for the Semantic Web', *IEEE Intelligent Systems*, **16**(2), 72-79, (2001).

[19] Z.S. Harris ZS, *Mathematical Structures of Language*, Wiley, New York, 1968.

[20] E. Agirre et al., 'Enriching very large ontologies using the WWW', *ECAI Workshop on Ontology Learning,* (2000).

[21] A. Faatz, R. Steinmetz, 'Ontology Enrichment with Texts from the WWW', *Machine learning: proceedings/ ECML 2002, 13th European Conference on Machine Learning,* Helsinki, Finland, (2002).

[22] S. Kullback, R.A. Leibler, 'On Information and Sufficiency', *Ann. Math. Statis*. **22**(1), 79-86, (1951).

[23] D. Sanchez, A. Moreno, 'Creating Ontologies form Web Documents', *Recent Advances in Artificial Intelligence Research and Development,* **113**, 11-18, (2004).

[24] P. Cimiano, A. Hotho, S. Staab, 'Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis', *Journal of Artifical Intelligence Research,* **24**, 305-339, (2005).

[25] H. Schütze, *Foundations of statistical natural language processing*, The MIT Press., Massachusetts, 1999.

[26] U. Hahn, KG Marko, 'Joint Knowledge Capture for Grammars and ontologies', *K-CAPI'01 Conference*, 68-76, (2001).

[27] M. Kietz, R. Volz, and A. Mädche, 'A method for semi-automatic ontology acquisition from a corporate intranet', *Proceedings of CoNLL-2000 and LLL-2000*, Lisbon, 167-175, (2000).

[28] K.M. Gupta et al., 'An Architecture for Engineering Sublanguages WordNets', *The First International Conference on Global WordNet*, 21-25, (2002).

[29] E. Alfonseca and S. Manandhar, 'Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures', *EKAW-2002, 1–7,* (2002).

[30] S. Narr and E.W. DeLuca and S. Albayrak, 'Extracting semantic annotations from Twitter', *ESAIR'11*, 15-16, (2011).

Annex

Ontology Learning Approach	Description
Agirre et. al (2000)	*‘Enriching very large ontologies using the WWW’;* Statistical ontology learning approach; [20]
Faatz and Steinmetz (2002)	*’Ontology Enrichment with Texts from the WWW’,* Statistical ontology learning approach; [21]
Sanchez and Moreno (2004)	*‘Creating Ontologies form Web Documents’;* Statistical ontology learning approach; [23]
Cimiano, Hotho and Staab (2005)	*‘Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis’;* Statistical ontology learning approach; [24]
Hearst (1992)	*‘Automatic Acquisition of Hyponyms from Large Text Corpora’;* NLP approach; [15]
Kietz, Volz and Mädche (2000)	*‘A method for semi-automatic ontology acquisition from a corporate intranet’;* NLP approach; [27]
Gupta et al. (2002)	*‘An Architecture for Engineering Sublanguages WordNets’;* NLP approach; [28]
Alfonseca and Manandhar (2002)	*‘Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures’;* NLP approach; [29]
Narr, DeLuca, Albayrak (2011)	*‘Extracting semantic annotations from Twitter’;* NLP approach; [30]

Table 1: Overview of Ontology Learning Approaches

Criterion	Description
Criterion 1: Goal	Detection and analysis of the specific problem which the ontology learning approach is dealing with.
Criterion 2: Methodology	Detection and analysis of the underlying methodological approach within statistical or NLP approaches for ontology learning.
Criterion 3: Technique	Detection and analysis of the technical approach of the ontology learning methodology. This criterion is a specification of criterion 2 and provides more technical information.
Criterion 4: Degree of Automation	Detection of the grade of automation with which the respective ontology learning approach is able to perform in a (partly) unsupervised way.
Criterion 5: Ontology Reuse	Detection whether the ontology learning approach uses already existing ontologies or other bodies of knowledge.
Criterion 6: Ontology	Detection and analysis of the derived ontology elements and possible conceptual structures.

Table 2: Overview of Comparison Criteria

	Methodology; Goal	Technique; Derived Ontology	Grade of Automation; Ontology Reuse
[20]	overcoming word ambiguity; establish relationships between words	Harris‘ distributional hypothesis as basic methodology, topic signatures to measure relevance of retrieved data, clustering by statistical measures, ontology is enriched with concepts and relationships	automatic, but with biased results (unsupervised); reuse of WordNet
[21]	enriching existing ontology	predefined text resources as starting point, clustering and statistical information, similarity measures; nouns and relationships	semi-automatic, domain experts select data sources; use of medical ontologies
[23]	creating new ontology for a specific domain; enriching ontology	Extracting information by key words; Classes	automatic; none
[24]	creating ontologies	Formal Concept Analysis, based on Harris‘ distributional hypothesis; concepts and taxonomies	semi-automatic, domain experts select the data sources; none
[15]	enriching existing ontologies	retrieving lexico-syntactic patterns; nouns, hyponym-hypernym relationships	automatic; reuse of WordNet
[27]	creating an ontology-centered application on the basis an existing ontology	combines statistical and pattern-based techniques; is-a relationships, concepts	semi-automatic; reuse of GermaNet, WordNet
[28]	creating sublanguage WordNets to enrich an upper ontology	Retrieval of sublanguage WordNets, adding synsets, updating mechanism; WordNets	semi-automatic, domain experts select data sources; reuse of WordNet
[29]	enriching ontology with domain specific information	unsupervised methodology derives concepts from domain specific data sources, integration in existing lexical ontologies; concepts	automatic; reuse of WordNet
[30]	enriching existing ontologies	ontologies are derived from the social media service Twitter; annotations, contextual relationships.	automatic; not explicitly reported

Table 3: Results of Comparing Ontology Learning Approaches

Description Logic Reasoning in an Ontology-Based System for Citizen Safety in Urban Environment

Weronika T. Adrian, Jarosław Waliszko, Antoni Ligęza, Grzegorz J. Nalepa, Krzysztof Kaczor[1]

Abstract. Semantic annotations and formally grounded ontologies are flexible yet powerful methods of knowledge representation. Using them in a system allows to perform automated reasoning and can enhance the knowledge management. In the paper, we present a system for collaborative knowledge management, in which an ontology and ontological reasoning is used. The main objective of the application is to provide information for citizens about threats in an urban environment. The system integrates a database and an ontology for storing and inferring desired information. While a terminology of the traffic danger domain is described by the ontology, the location details of traffic conditions are stored in the database. During run-time, the ontology is populated with instances stored in the database and used by a Description Logic reasoner to infer required facts.

1 Introduction

One of the important research fields of Artificial Intelligence (AI) is the area of Knowledge Representation and Reasoning (KR&R) [3]. The Semantic Web [2] initiative is sometimes perceived as the new incarnation of AI, tackling some of its problems and challenges. Although this worldwide project is not aimed at constructing intelligent machines, it has resulted in development of several effective KR&R methods. Representation of knowledge is done on a few levels of abstraction. For concrete objects, attributes and relations to other objects (resources) are defined by use of semantic annotations organized into semantic vocabularies for various domains. Classification of objects and classes definition using their interdependencies is done with use of ontologies [4] of different expressiveness and formality level. Stating logical axioms about classes enable automated reasoning and inferring conclusions about concrete objects. In the end, semantic applications can make use of this multilevel knowledge representation and exhibit semi-intelligent behavior.

Web-based information systems have been widely used to facilitate communication and distribution of information in a rapid and efficient way. Whether through official news portals or social systems like Facebook or Twitter, people inform each other about the events or dangers. Using GIS systems [11] that allow to store, represent and search geographic information, users can add location metadata to the information they provide or get useful data based on their localization (e.g. by the use of a GPS). Projects such as Wikipedia has demonstrated that people are willing to cooperate if they find it worthwhile and the system is easy to use. Collaborative knowledge engineering and management can be enhanced by employing intelligent techniques, for example by using underlying knowledge representation. However, a system interface must remain simple.

We propose a system for collaborative knowledge management enhanced with semantic knowledge representation and reasoning. The main objective of the system is to gather knowledge about threats of various sorts within a defined urban area. The system should serve the local community and the police. Our proposed solution combines social software features (commenting, ratings etc.) with a strong underlying logical representation of knowledge entered by users. The application employs AI methods, namely a domain ontology of traffic dangers and conditions and a Description Logic (DL) [1] reasoner to infer knowledge from facts explicitly present in the system.

The rest of the paper is organized as follows: in Section 2 the motivation for our research is given with references to selected previous works. Section 3 gives an overview of the system, including its functionality, architecture, a threat ontology, the integration of an ontology and a database in the system, the reasoning in the system and the user interface. The implementation is briefly described in Section 4. Overview of related works is presented in Section 5. The paper is summarized in Section 6 and future work is outlined in Section 7.

2 Motivation

Within the INDECT Project [2] important problems related to security and intelligent information systems are investigated. Task 4.6 of the project focuses on development of a *Web System for citizen provided information, automatic knowledge extraction, knowledge management and GIS integration* [6]. The main objective of our research is to develop a semantically enriched environment for collaborative knowledge management. Local communities should be able to quickly share information about current traffic dangers and threats, for instance closed roads, holes in the pavements and streets, dangerous districts or events that impede a normal traffic. The system proposed within the task should be a community portal that allows citizens to participate and cooperate in order to improve the security in the urban environment. Within the task several initial system prototypes have been developed [5, 10] and the current work consists in integrating the best solutions for the final system.

The system should use some sort of intelligent processing to provide possibly most useful knowledge to the users. To this end, a Knowledge-Based System (KBS) should be proposed, with a formalized knowledge representation and reasoning mechanisms. Categorization of threats and possibility of inferring new facts based on the ones entered by users is a desired feature. To enhance the automated knowledge processing of the system, semantic technologies for GIS were analyzed and discussed in [7].

While a threat domain ontology can be the same for various locations, different system installations will vary depending on the lo-

[1] AGH University of Science and Technology, Poland, email: {wta,ligeza}@agh.edu.pl

[2] See `http://indect-project.eu`.

cations they work in. Abstract knowledge can and should be shared across applications boundaries to facilitate change management. On the other hand, the system should be robust and easily adaptable to local conditions, so the access to the actual data should be optimized.

The system should encompass social features, such as possibility to comment on, discuss and rate information entered by other users. This way, the users can gain or loose credibility and the community can indirectly control spam information. The user interface (UI) should be intuitive and easy to use, potentially adaptable to various hardware platforms including desktop and mobiles. Encompassing these requirement should provide a useful intelligent system for improving urban safety.

3 System Overview

The proposed system is an ontology-driven application. It integrates a database and an ontology for storing and inferring knowledge about traffic dangers in a given area. While the abstract of traffic danger domain is described by the ontology, the location details of traffic conditions and geographical information (e.g. relations among concrete streets, districts and postal codes) are stored in a relational database. During the run-time, the information from the database is integrated (synchronized) with the core ontology (the terminology is populated with instance data stored in a database). The synchronization is done automatically at the application start and at any time on a user demand. The synchronized ontology is then used by a DL reasoning engine to infer facts about chosen area. The deduction is based on definitions of threats which depends on specific traffic conditions present in specific locations.

3.1 Functionality and User Roles

The main objective of the proposed system is to provide citizens with a real-time data about dangers occurring in a chosen area. Part of this information is entered into the system by so-called *trusted users*. The rest is automatically inferred based on the axioms of the threat ontology and instance data of current conditions (facts).

Three kinds of users are distinguished within the system. *Regular users* can browse the system knowledge and ask questions about specific locations and dangers. They address the system with dynamic questions and get results of inferred traffic dangers. *Trusted users* can modify the information stored in the database, e.g. they can update the locations of traffic conditions occurrences. The information is validated and stored in the database. Updated knowledge can be used for dangers deduction process, after *synchronization* with the ontology. Finally, the *experts* can modify the core ontology.

3.2 Architecture and Data Flow

The system is divided into three functionally different layers:

- a web dashboard layer dedicated to the interaction with users (through browser clients),
- a platform layer which is the core of system responsible for processing knowledge, and
- a storage layer, where all the data is stored, in a database and an ontology (see Figure 1).

All users can interact with the web-based dashboard for querying system, to get desired information. The main logic of the system (presented in Fig. 1 as three cogged wheels) consists in: downloading the core ontology (*"Getting core ontology from URI"*), synchronizing the core ontology with the current data uploaded by trusted users (*"Synchronization with database"*), and inferring the ontology dependencies (*"Knowledge inferencing"*).

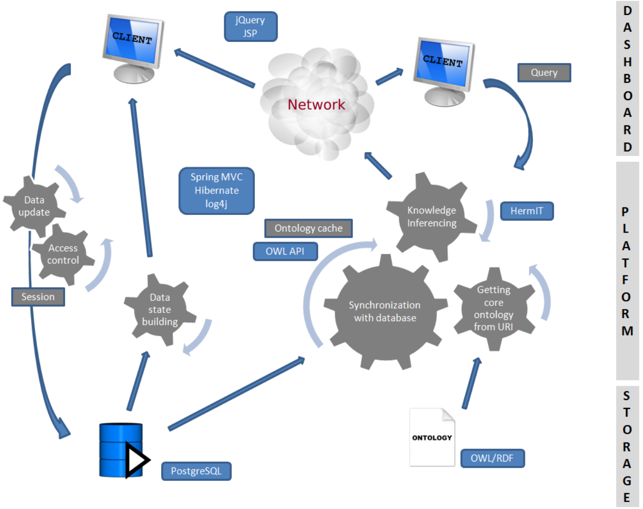


Figure 1. Data flow in the system

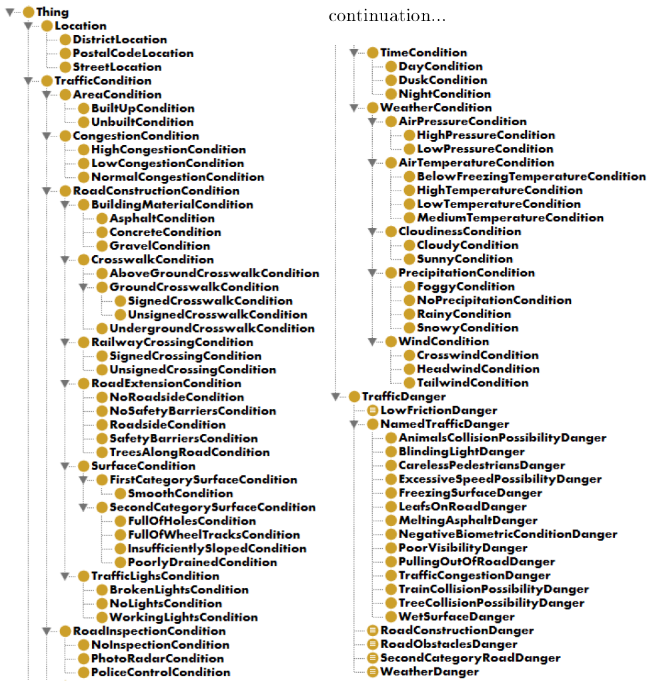
For working with most recent data, provided by trusted users, the *synchronization* mechanism integrates core ontology, describing the abstract of traffic dangers, with specific real time data. The process is executed for the first time on application start, i.e. the first request to the server while accessing the main page of the system. This functionality is also available on demand. After synchronization, the populated ontology is cached in memory and used for inferencing.

A single installation of the system (for instance for a single city) has its own database, in which the information about streets, districticts and actual conditions are stored. The *core ontology* on the other had can be shared by several installations of the system. It is accessible by an URI and can be stored on local or remote server.

3.3 Traffic Danger Ontology

As noted in [8], "In recent years the development of ontologies has been moving from the realm of Artificial-Intelligence laboratories to the desktops of domain experts.". Sharing common understanding of the knowledge domain is one of the most critical reason for developing ontologies. Explicit domain assumptions provide a clear description of the domain knowledge and simplify the knowledge extensibility. In our case, the domain ontology consists of concepts of geographical locations (streets, districts, postal codes), traffic dangers (e.g. *LowFrictionDanger*, *RoadConstructionDanger*), traffic conditions (including among others a hierarchy of *WeatherCondition*s and *RoadConstructionCondition*s) and describes multiple relations among them. An excerpt of the ontology is shown in Figure 2. The ontology enable the system to reason upon stored facts and answer the questions of the following types:

- What traffic dangers can be encountered within a specific area?
- Is there any danger within area of specific postal code or specific district?
- What kind of dangers are connected with specific atmospheric conditions?
- Are there any dangers connected with specific condition (e.g. low friction) in a specific area?



- Thing
  - Location
    - DistrictLocation
    - PostalCodeLocation
    - StreetLocation
  - TrafficCondition
    - AreaCondition
      - BuiltUpCondition
      - UnbuiltCondition
    - CongestionCondition
      - HighCongestionCondition
      - LowCongestionCondition
      - NormalCongestionCondition
    - RoadConstructionCondition
      - BuildingMaterialCondition
        - AsphaltCondition
        - ConcreteCondition
        - GravelCondition
      - CrosswalkCondition
        - AboveGroundCrosswalkCondition
        - GroundCrosswalkCondition
          - SignedCrosswalkCondition
          - UnsignedCrosswalkCondition
        - UndergroundCrosswalkCondition
      - RailwayCrossingCondition
        - SignedCrossingCondition
        - UnsignedCrossingCondition
      - RoadExtensionCondition
        - NoRoadsideCondition
        - NoSafetyBarriersCondition
        - RoadsideCondition
        - SafetyBarriersCondition
        - TreesAlongRoadCondition
      - SurfaceCondition
        - FirstCategorySurfaceCondition
          - SmoothCondition
        - SecondCategorySurfaceCondition
          - FullOfHolesCondition
          - FullOfWheelTracksCondition
          - InsufficientlySlopedCondition
          - PoorlyDrainedCondition
      - TrafficLighsCondition
        - BrokenLightsCondition
        - NoLightsCondition
        - WorkingLightsCondition
    - RoadInspectionCondition
      - NoInspectionCondition
      - PhotoRadarCondition
      - PoliceControlCondition

continuation...

-
  -
    - TimeCondition
      - DayCondition
      - DuskCondition
      - NightCondition
    - WeatherCondition
      - AirPressureCondition
        - HighPressureCondition
        - LowPressureCondition
      - AirTemperatureCondition
        - BelowFreezingTemperatureCondition
        - HighTemperatureCondition
        - LowTemperatureCondition
        - MediumTemperatureCondition
      - CloudinessCondition
        - CloudyCondition
        - SunnyCondition
      - PrecipitationCondition
        - FoggyCondition
        - NoPrecipitationCondition
        - RainyCondition
        - SnowyCondition
      - WindCondition
        - CrosswindCondition
        - HeadwindCondition
        - TailwindCondition
  - TrafficDanger
    - LowFrictionDanger
    - NamedTrafficDanger
      - AnimalsCollisionPossibilityDanger
      - BlindingLightDanger
      - CarelessPedestriansDanger
      - ExcessiveSpeedPossibilityDanger
      - FreezingSurfaceDanger
      - LeafsOnRoadDanger
      - MeltingAsphaltDanger
      - NegativeBiometricConditionDanger
      - PoorVisibilityDanger
      - PullingOutOfRoadDanger
      - TrafficCongestionDanger
      - TrainCollisionPossibilityDanger
      - TreeCollisionPossibilityDanger
      - WetSurfaceDanger
    - RoadConstructionDanger
    - RoadObstaclesDanger
    - SecondCategoryRoadDanger
    - WeatherDanger

Figure 2. Traffic Danger Ontology: Asserted class hierarchy

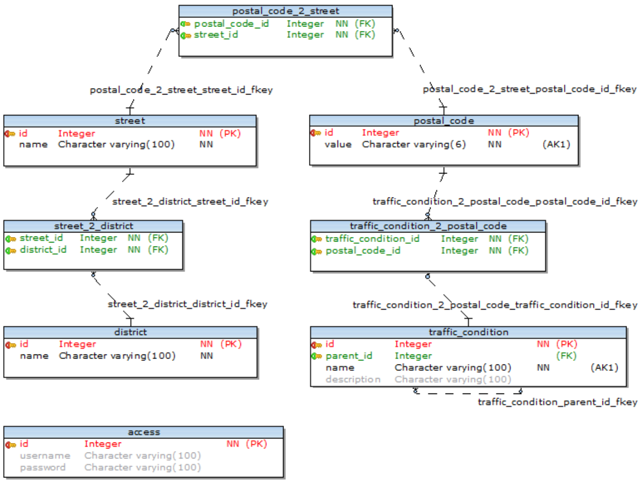
- What are the sub-areas of a specific location?
- Are there any traffic conditions provided for a specific location?

In order to answer these questions either a semantic reasoner is used (which performs classification of concepts within the ontology) or appropriate DL queries are constructed. Based on the definitions of *TrafficDanger*s in the ontology and information about actual condition occurences, implicit knowledge may be deduced (what kind of danger results from given conditions in a selected area).

3.4 Integration of the Ontology and the Database

While the abstract domain knowledge is expressed by the ontology axioms, the operational knowledge of the system is stored in a relational database. The database schema can be observed in Figure 3. The knowledge stored in the database consists of the locations structure and the actual traffic conditions in these locations. Specifically, the locations of the traffic conditions occurrences are defined by postal codes. The postal codes are connected to streets, which in turn are connected to districts. For instance, one can add an information that a particular street is under construction (a *RoadConstructionCondition* or one of its subclasses occurs) or that there is a specific weather condition in a specific district.

One of the most important aspects of the system is the possibility of an integration of data from the database and the ontology. Upon the *synchronization* process, the core ontology is cached and populated with the data from the database becoming a *synchronized ontology*. While the core ontology describes a terminology of traffic danger, the synchronized one is related to a specific environment and used for reasoning. Consequently, synchronized ontology can differ between the various environments where the system is deployed. For example, traffic conditions information for Cracow can vary significantly from those in Montpelier. Although it is possible to have a single installation of the system and synchronizing the ontology at once with all global data, it can result in system overloading and decreasing performance while inferring dependencies.



postal_code_2_street			
postal_code_id	Integer	NN	(FK)
street_id	Integer	NN	(FK)

street			
id	Integer	NN	(PK)
name	Character varying(100)	NN	

postal_code				
id	Integer	NN	(PK)	
value	Character varying(6)	NN		(AK1)

street_2_district			
street_id	Integer	NN	(FK)
district_id	Integer	NN	(FK)

traffic_condition_2_postal_code			
traffic_condition_id	Integer	NN	(FK)
postal_code_id	Integer	NN	(FK)

district			
id	Integer	NN	(PK)
name	Character varying(100)	NN	

traffic_condition				
id	Integer	NN	(PK)	
parent_id	Integer		(FK)	
name	Character varying(100)	NN		(AK1)
description	Character varying(100)			

access			
id	Integer	NN	(PK)
username	Character varying(100)		
password	Character varying(100)		

postal_code_2_street_street_id_fkey
postal_code_2_street_postal_code_id_fkey
street_2_district_street_id_fkey
traffic_condition_2_postal_code_postal_code_id_fkey
street_2_district_district_id_fkey
traffic_condition_2_postal_code_traffic_condition_id_fkey
traffic_condition_parent_id_fkey

Figure 3. ER diagram of traffic database

3.5 Reasoning in the System

Reasoning in the system is provided by invoking a DL reasoner on a synchronized ontology. The sequence diagram (see Figure 4) the required steps for the reasoning process. Once the trusted users have provided traffic conditions facts, a regular user can check what threat they may expect in a specific area. Responding to the user request, the system imports the up-to-date facts into a locally stored ontology (synchronizes the ontology), and then query the ontology by posing appropriate DL queries. From a user perspective, a query is constructed by selecting a desired location through a web-based interface. Once the query is created, the DL reasoner is invoked to process it on the cached ontology. The inferred set of information is provided to the user. The reasoning is time-efficient although no formal tests has been done yet. This is a subject for further work.

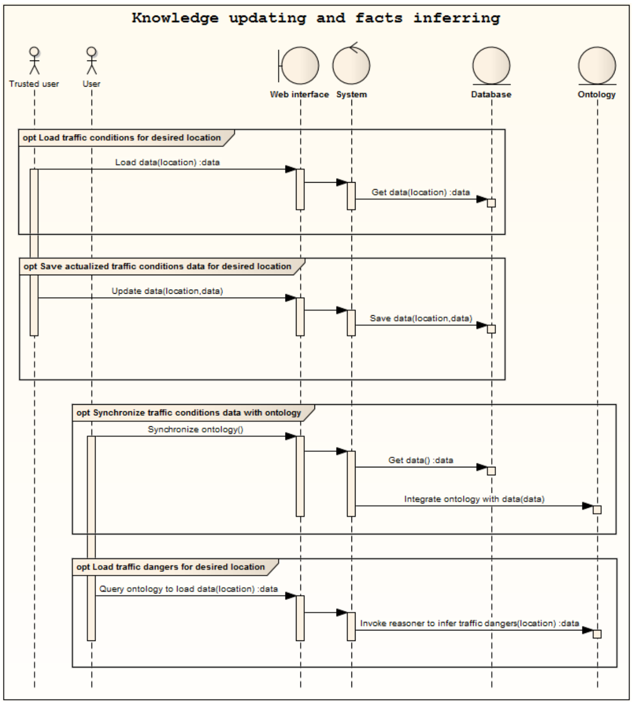
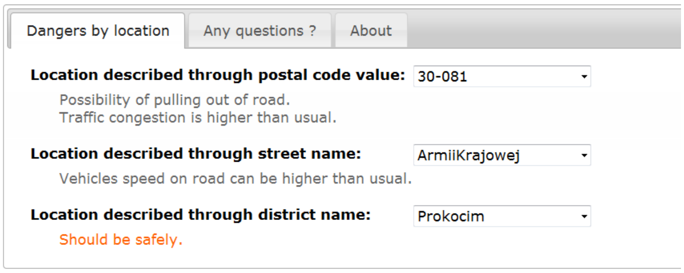

Knowledge updating and facts inferring
Trusted user
User
Web interface
System
Database
Ontology
opt Load traffic conditions for desired location
Load data(location) :data
Get data(location) :data
opt Save actualized traffic conditions data for desired location
Update data(location,data)
Save data(location,data)
opt Synchronize traffic conditions data with ontology
Synchronize ontology()
Get data() :data
Integrate ontology with data(data)
opt Load traffic dangers for desired location
Query ontology to load data(location) :data
Invoke reasoner to infer traffic dangers(location) :data

Figure 4. Sequence diagram for updating and inferring data

3.6 User Interface

The web-based interface of the system allows its users to create dynamic questions, and get the results about inferred traffic dangers. The prototype implementation [9] uses simple forms by use of which the users can construct questions for the system, e.g. a user can choose a desired location from a drop-down list and ask what threats may be encountered in this particular area (see Fig. 5).



Location described through postal code value: 30-081

Possibility of pulling out of road.
Traffic congestion is higher than usual.

Location described through street name: ArmiiKrajowej

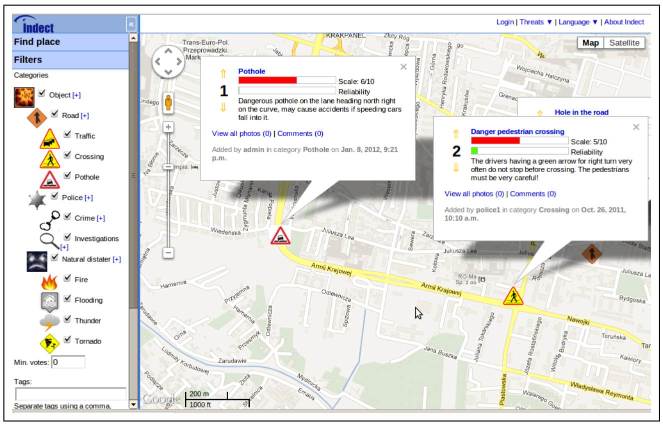
Vehicles speed on road can be higher than usual.

Location described through district name: Prokocim

Should be safely.

Figure 5. An excerpt of web-based user interface of the system

As this is a position paper, the full development of a Graphical User Interface (GUI) with a map component is still in progress. There is a process going on to integrate the logical layer of the system with an interface that uses maps and provides social features for the community of users. A fully-fledged GUI with an interactive map on which the users can navigate and filter threats by location, date, severity etc. has been developed (see Fig. 6), but is not yet fully integrated with the logical layer. In this version, a spatially-enabled database will be used which allows to store geographical data in an efficient way. The usability of the system is expected to increase, due to the possibility of visually choosing an area of interest (see Fig. 7) and the social features like possibility to rate threats and discuss them.



Pothole
1
Scale: 6/10
Reliability
Dangerous pothole on the lane heading north right on the curve, may cause accidents if speeding cars fall into it.
View all photos (0) | Comments (0)
Added by admin in category Pothole on Jan. 8, 2012, 9:21 p.m.

Danger pedestrian crossing
2
Scale: 5/10
Reliability
The drivers having a green arrow for right turn very often do not stop before crossing. The pedestrians must be very careful!
View all photos (0) | Comments (0)
Added by police1 in category Crossing on Oct. 26, 2011, 10:10 a.m.

Figure 6. GUI for the system with an interactive map and search filters.

4 Implementation and Deployment

The ontology has been developed in a top-down process with the Protégé [3] editor integrated with the HermiT DL Reasoner [4]. The ontology is provided in different formats (OWL2 XML [5], RDF/XML [6] or OWL2 Manchester Syntax) [7]). The synchronization is based on the OWL API library [8] and provides up-to-date information (cache in memory) for the HermiT DL Reasoner.

The ontology can be stored on local or remote server and is accessed by an URI. The cooperation with a database is provided through the Hibernate ORM [9] technology. The simple form-based user interface has been built with the JavaServer Pages (JSP) [10] and jQuery JavaScript Library [11], while requests from users and appropriate responses, are controlled by Spring MVC [12]. For logging the results of particular operations, log4j Java-based logging utility [13] is used. PostgreSQL [14] is choosen as SQL database. The application has been written in Java using the Eclipse Java IDE [15]. Dependencies management and versioning is the task of Apache Maven tool [16]. All these technologies are free software or open source.

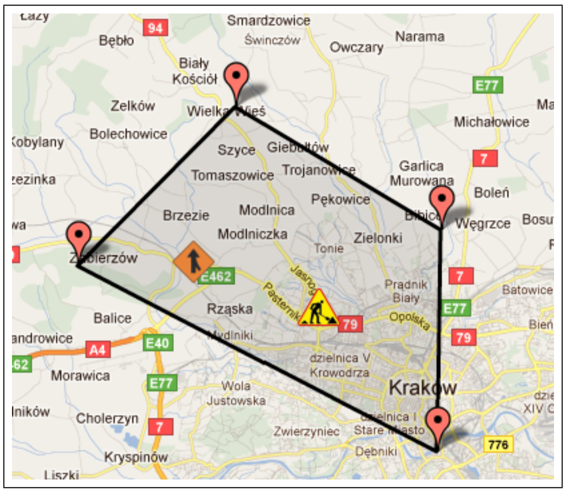


Figure 7. GUI for the system: a map with selected area.

5 Related Work

Crime Mapping systems were originally a class of systems that map, visualize and analyze crime incident patterns using Geographic Information Systems (GIS). This name has been later extended to incorporate all applications that aid in improving the public safety. This include natural disasters monitoring systems which are often designed for specific regions and the scope of their functionalities is usually limited to the specific types of disasters that are most common and most dangerous in those regions, systems monitoring threats on the roads and crime monitoring systems. A detailed survey of the existing crime mapping systems is given in [12]. Here, only a few of them are mentioned.

Examples of systems monitoring threats on roads are: Traffic Information Highways Agency (reporting bad weather, speed reductions, accidents and road works in England, `http://`

[3] See `http://protege.stanford.edu/`
[4] See `http://owlapi.sourceforge.net/`
[5] See `http://www.w3.org/TR/owl2-xml-serialization/`
[6] See `http://www.w3.org/TR/rdf-syntax-grammar/`
[7] See `http://www.w3.org/TR/owl2-manchester-syntax/`
[8] See `http://owlapi.sourceforge.net/.`
[9] See `http://www.hibernate.org/`
[10] See `http://www.oracle.com/technetwork/java/javaee/jsp/index.html`
[11] See `http://jquery.com/`
[12] See `http://static.springsource.org/spring/docs/3.0.x/`
[13] See `http://logging.apache.org/log4j/`
[14] See `http://www.postgresql.org/`
[15] See `http://www.eclipse.org/`
[16] See `http://maven.apache.org/`

www.highways.gov.uk/traffic/traffic.aspx), Travel News Scotland (accidents, road works, other difficulties, http://www.bbc.co.uk/travelnews/scotland) or TravelMidwest.com (road capacity, road works, accidents, speed cameras, etc. in Chicago and several other cities and areas around Midwestern states, http://www.travelmidwest.com). NAVTEQ (http://www.traffic.com/) is one of the few of such services that attempt to provide traffic information from all around the world, yet it only provides information about capacity and delays, not about threats or accidents. In most cases, the represented data is collected automatically by sensors and readers, such as inductive loops and automatic number plate recognition cameras and then processed centrally. For instance, in case of Traffic England (http://www.trafficengland.com), TrafficScotland (http://trafficscotland.org/) and TrafficWales (http://www.traffic-wales.com), this information is compiled by Highway Agency's National Traffic Control Centre.

The systems present the threats in a visual form on a map and provide various output channels, e.g. e-mail notifications, text messages or even Twitter alerts. They operate on mobile devices and make use of their GPS systems. The apparent lack, however, is that the information presented to the users is strictly that which was entered. The original contribution of our approach, although still in a development phase, is to supply the system with intelligent processing techniques based on ontological reasoning. Moreover, our approach aims at encompassing various kinds of threats by using a threat ontology.

6 Summary

AI techniques may be successfully used in various applications for Knowledge Management. Using an ontology in a KM system allows to store abstract data, share it across several installations and manage changes in a centralized way. A loose coupling of the ontology with a relational database allows to store concrete data about conceived area in a database and populate the ontology with instance data during application run-time. Embedding a Description Logics reasoner enable the system to reason upon explicit knowledge entered by users and give back a useful response. A graphical user interface with a map component and social software features make the system user friendly and has a gradual learning curve.

To the best of our knowledge, there does not exist a crime mapping system that uses ontologies and DL reasoning to provide rich information based on knowledge gathered in the system. Although there exist numerous solutions for various danger information systems, none of them describe the threats in a formalized ontological way, relate weather or road conditions to the possible dangers and reasons about these dependencies. We believe that this is our original contribution compared to existing literature.

7 Future work

The system has been tested with several Web browsers and can be used on any device that support Web browsing. However, for mobile devices, some adaptations are needed. The current prototype implementation has a limited user interface. The intended integration with a GUI providing interactive map and social features is not yet finalized. A possible direction for further development could be focused on extensions for heterogeneous application-to-application communication. The RESTful Web Services [17] can be considered. These external systems would be perceived as software agents. Their tasks could be focused on periodic connections to the system, getting some information set, and creating statistics about the traffic dangers. The statistics could visualize frequencies of particular dangers on a specific area or classify the safety of the selected district.

[17] See http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.

ACKNOWLEDGEMENTS

The research presented in this paper is carried out within the EU FP7 INDECT Project: "Intelligent information system supporting observation, searching and detection for security of citizens in urban environment" (http://indect-project.eu).

References

[1] *The Description Logic Handbook: Theory, Implementation, and Applications*, eds., Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, Cambridge University Press, 2003.

[2] Tim Berners-Lee, James Hendler, and Ora Lassila, 'The Semantic Web', *Scientific American*, (May 2001).

[3] Ronald Brachman and Hector Levesque, *Knowledge Representation and Reasoning*, Morgan Kaufmann, 2004.

[4] Nicola Guarino, 'Formal ontology and information systems', in *Proceedings of the First International Conference on Formal Ontologies in Information Systems*, pp. 3–15, (1998).

[5] Antoni Ligęza, Weronika T. Adrian, Sebastian Ernst, Grzegorz J. Nalepa, Marcin Szpyrka, Michał Czapko, Paweł Grzesiak, and Marcin Krzych, 'Prototypes of a web system for citizen provided information, automatic knowledge extraction, knowledge management and gis integration', in *Multimedia Communications, Services and Security*, eds., Andrzej Dziech and Andrzej Czyżewski, volume 149 of *Communications in Computer and Information Science*, 268–276, Springer Berlin Heidelberg, (2011).

[6] Antoni Ligęza, Sebastian Ernst, Sławomir Nowaczyk, Grzegorz J. Nalepa, Weronika T. Furmańska, Michał Czapko, Paweł Grzesiak, Marcin Kałuża, and Marcin Krzych, 'Towards enregistration of threats in urban environments : practical consideration for a GIS-enabled web knowledge acquisition system', in *MCSS 2010 : Multimedia Communications, Services and Security : IEEE International Conference : Kraków, 6-7 May 2010*, ed., Andrzej Głowacz Jacek Dańda, Jan Derkacz, pp. 152–158, (2010).

[7] Grzegorz J. Nalepa and Weronika T. Furmańska, 'Review of semantic web technologies for GIS', *Automatyka: półrocznik Akademii Górniczo-Hutniczej im. Stanisława Staszica w Krakowie*, **13**(2), 485–492, (2009).

[8] Natalya F. Noy and Deborah L. McGuinness, *Ontology Development 101: A Guide to Creating Your First Ontology*, Stanford University, Stanford, CA, 94305.

[9] Jarosław Waliszko, *Knowledge Representation and Processing Methods in Semantic Web*, Master's thesis, AGH University of Science and Technology, 2010.

[10] Jarosław Waliszko, Weronika T. Adrian, and Antoni Ligęza, 'Traffic danger ontology for citizen safety web system', in *Multimedia Communications, Services and Security*, eds., Andrzej Dziech and Andrzej Czyżewski, volume 149 of *Communications in Computer and Information Science*, 165–173, Springer Berlin Heidelberg, (2011).

[11] *The Handbook of Geographic Information Science*, eds., John P. Wilson and A. Stewart Fotheringham, Blackwell Publishin Ltd, 2008.

[12] Maciej Żywioł. Analysis and evaluation of crime mapping systems, 2012.

Advanced System for Acquisition and Knowledge Management in Cultural Heritage.

Stefan du Château and **Danielle Boulanger** and **Eunika Mercier-Laurent** [1]

Abstract. In this paper we present our research work on the system of acquisition and knowledge management in cultural heritage. This is a hybrid system because it uses several techniques, signal and natural language processing and knowledge modelling to effectively help a researcher in cultural patrimony in collecting, recording and finding the relevant knowledge.

- The signal processing is used for the oral description of the cultural heritage and for transcription from voice to text.
- The linguistic analysis is used for search and extraction of information.
- Modelling of knowledge is used for the creation of a partial ontology of domain according to a model predefines.

After introducing the problem of on field information collecting and managing, we describe the specific work of a researcher in the field of cultural heritage and main difficulties. Furthermore we explain our choice of the architecture of this hybrid system, our experiments and the results. Finally we give some perspective on extending this system to the other domains.

1 Introduction

A common problem in knowledge engineering is the efficient collection of information and knowledge from sources considered to be scientifically reliable. These can be human experts, written records or computer applications (databases) that cover the domain knowledge. Depending on the situation, treatment and expected outcome, different collection methods can be used.

The work of researchers in the area of cultural heritage consists in one part of gathering of information in the field, in towns and villages in the form of text files, photos, sketches, maps and videos. If necessary, the information gathered for each work is corrected, archived, and finally stored in a database. The storage of information in paper documents or directly on laptops is cumbersome and time consuming. The amount of information collected is very large, the data is heterogeneous and its transformation into a form that can be used for research is not automatic.

The system we propose [1], uses a voice interface that reduces the amount of time used in the process of collection, because the description of the artefacts studied can be voice recorded and saved as an audio file. This is a hybrid system because it relies on technologies of signal processing, knowledge modelling and natural language processing.

2 Nature of knowledge in the field of cultural heritage

Knowledge of cultural heritage, is incremental, highly context dependent and multidisciplinary, more than in other disciplines.

- Incremental, because it cannot evolve without the prior knowledge.
- Context-dependent, because the creation and existence of an cultural heritage object is rarely random. Also its location in one place is often the result of intention. The history of an object can only be understood through the study of its context.
- Multidisciplinary, because the study of tangible and intangible aspects of an object requires contribution of other sciences such as chemistry, geology, archaeology, ethnology (…).

According to A. Iacovella, a historical object "is full of meaning" [2].

The full perception of this meaning depends on both: the related knowledge and its context.

We are thus dealing with a flow of knowledge including researchers in cultural heritage, associated domains experts, their knowledge and experiences, knowledge of related disciplines, constraints of existing Information System (Inventory Descriptive System, databases, lexicons ...); conceptual models, and interactions between these elements have also be considered. Capturing knowledge in a process of study and analysis of historical patrimony requires system, global and holistic thinking ability. The system approach is necessary for understanding of interactions between the information emerging from the studied object, knowledge of researcher studying this object, knowledge and information providing by the object context and those of other related sciences.

The information on a given object in its particular context can be understood only through the knowledge accumulated by a researcher and knowledge of other sciences at its disposal, so the “global thinking” is useful.

Finally, the knowledge about the considered object involves not only its intrinsic properties, shape, structure, dimensions, but depends also on its various contexts, it can be found - a “holistic thinking” may help [3][4].

[1] MODEME, Research Center IAE University of Jean Moulin 6, av Albert Thomas F-69008 Lyon France

2.1 Events

The existence of an object made by man is highly correlated with the notions of time and space. Indeed, the creation event took place on a given date or period and in a specific place.

During its existence, the object can undergo numerous other events such as moves, modification, repair or destruction that may result in changes of shape, materials, changes in the functions of the object and of involved actors (restaurateurs, producers, curators, etc.). All these events must be located in time and space. An event takes place in a given context.

2.2 Importance of context

In the field of historical sciences and especially in the cultural heritage area, the perception of the meaning of data and information on a studied object is strongly influenced by its context. Interpreting information about a physical object depends on the place it is. For example an iron host, shuttle incense, blessed bread basket found and studied outside its usual context (a church or sacristy), may lose some of the attributes can be called functional. Thus an iron host can be considered as an ordinary hammer, the blessed bread basket as a regular bread basket and the shuttle incense as sugar box. The context is even more significant in the field of archeology in which the interpretation of information and therefore the constitution of knowledge can not be done without taking into account the stratigraphic level and geographical areas, providing information on the chronology of the object and allowing a comparative study. The last can not be possible if we do not know the environment in which the object was found. To consolidate the knowledge about an artefact, we need to create a permanent to and from between the context-related knowledge, knowledge related to other objects found in the same place and information about the object studied. The importance of context is such that the interest of the study of an object can be challenged based on this one. What interest can have a ceramic shard alone in the middle of nowhere ?

Studying a historical object involves on the one hand manipulation of objective information on objects such as physical characteristics and, secondly, that of subjective information related to the context studied. In addition, as we have pointed out, the historical study is conditioned by the prior knowledge, which implies the incrementally of knowledge.

The nature of knowledge to process, as well as existent constraints and models library which can be reused implies the choice of models.

3 Architecture of our system

The architecture of our system takes into account several factors. First, it enables the implementation of three functional steps: the collection of information and knowledge in a specific context, information extraction and semi-automatic generation of a partial domain ontology supervised by a conceptual model. On the other hand, it must respect the constraints imposed by existing: the descriptive system of inventory, lexicons and thesauri and conceptual model CIDOC-CRM .

The process leading to the ontology of discourse of an object consists of several steps:

1. The voice acquisition of the description of a artefact
2. Transcription of audio file into a text file, using Dragon software that we have enriched with a specific vocabulary of cultural heritage.
3. Display the result text to allow expert correct it if errors
4. The linguistic analysis and information extraction [5], [6]. This stage leans on the XIP (Xerox Incremental Parser) [7] software, which we enriched by semantic lexicons and grammatical rules, specific of the domain of the cultural heritage.
5. Validation of information got in the previous stage.
6. Generation of ontology of objects described during the first stage. It is the transfer of an implicit information contained in the SDI (Descriptive System of the Inventory) [8], defined by the Department of Heritage Inventory, to the explicit knowledge represented by the domain ontology of cultural heritage.

The architecture of our system is shown in the Figure1.

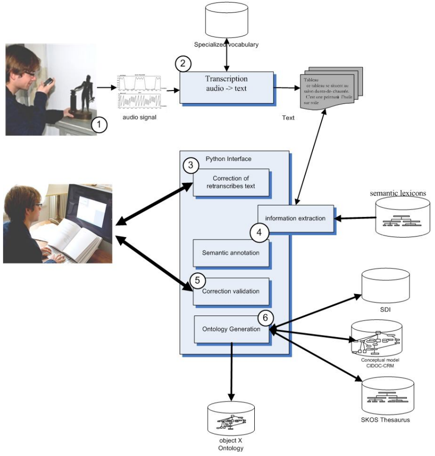


Figure 1. The architecture of Simplicius system

3.1 From voice to text

The audio file "translated" into text using the Dragon software; we have chosen it for its robustness and its performance in speech recognition.

Text files serve as information retrieval so that information is distributed into fields such as: NAME, CLASS, MATERIALS, DESCRIPTION, REGISTRATION (...), without requiring the speaker to specify the description field. The above fields derive from the descriptive system defined by the Department of Heritage Inventory. Some of these fields are mandatory, others optional. The content of certain fields is defined by a lexicon; the contents of other fields remain free.

Currently the data acquisition is done via keyboard and the user has to respect a highly structured data entry form. In the case of voice acquisition, there is no structure required to guide the user, who is usually a specialists in the field; we can therefore assume that the verbal description will be coherent and well structured. This has been proven in our experiments.

3.2 Analysis of resulting text and information extraction

To analyze the transcribed text (cf stage 4, figure 1), we use the robust XIP parser. This guarantees a result of corpus analysis, even if the text is malformed or erroneous, which can happen if the text is the result of an oral transcript [9], [10].

As we mention in Section 2.1, the information that has to be identified for extraction is defined by the above-mentioned descriptive system for artwork inventories, which defines not only the type of information that is to be looked for, but also controls, in some cases, the vocabulary to be used. The terms used should match the entry of a lexicon. The descriptive system of the inventory will therefore partially guide the creation of design patterns and of local grammars.

3.2.1 Lexicons

Two types of lexicons have been created: one that contains the vocabulary defined as authorized to fill out fields such as DENO, REPR MATR (...), and other which contains vocabularies for context analysis. Two types of formats are used. For lexicons with extended vocabularies, the term of each has been associated with its infinitive form for verbs and the masculine singular for nouns. In addition, its semantic and morphological trait was added to each term, as shown below:

calice
calice
+Denomination+Masc+Sg+Common+Noun
calices
calice
+Denomination+Masc+Pl+Common+Noun

The format of smaller glossaries includes the lemmatised form of the term and the semantic and morphological trait associated with it :

marque : noun += [!insc:+].
cachet : noun += [!insc:+].

3.2.2 Resolution of ambiguities

The identification of words or phrases is not the only difficulty faced by a system of information extraction. In the context-rich environment of cultural heritage artefact descriptions, the complexity of the language itself and the multiplicity of meanings that can be given to the descriptors used, one of the major problems is the resolution of semantic ambiguity. A word or phrase can be used in different contexts both to describe the characteristics of an artefact as well as the artefact itself, for example *a picture of a chalice,* the name of a person can be that of a person represented, or that of the artist (...). Often, heritage objects that are being described are part of a whole. The description of this type of object can refer to included elements, or to its container. It is therefore in a situation where several artefact names are mentioned. How do we know which is the subject of study ?

In the sentence: ***Calice** en argent doré, orné de grappes de raisins, d'épis de blé, de roseaux sur le pied et la fausse coupe, d'une **croix** et des **instruments** de la passion dans des **médaillons**, sur le pied.*

The terms: **calice**, **croix**, **instruments**, **médaillons** exist in the lexicon DENOMINATION. The term **calice** also exists in the lexicon REPRESENTATION

How can we be sure that, in this case, it is DENOMINATION?

How to choose the term for the DENOMINATION?

Study of the initial position

The study of the ordering of descriptors in a text provides valuable assistance, particularly for solving certain types of ambiguities. The study of the initial position, based on cognitive considerations [11], [12], gives special importance to the beginnings of sentences: the information at the beginning is a given information or at least one that is important.

In this perspective, extracting information from the following text:

***Calice** en argent doré, orné de grappes de raisins, d'épis de blé, de roseaux sur le **pied** et la fausse **coupe**, d'une **croix** et des instruments de la passion dans des **médaillons**, sur le pied.*

Will give a preference to the descriptor ***Calice*** compared to other descriptors mentioned above, to designate the name of the object studied.

Local context

Resolving ambiguities requires an analysis and understanding of local context. A morphosyntactic analysis of words surrounding the word whose meaning we seek to identify, as well as searching for linguistic clues in the context of a theme, can resolve some ambiguities.

In the sentence : *C'est une peinture à l'huile de très grande qualité, panneau sur bois **représentant** deux figures à mi corps sur fond de paysage**, Saint Guilhem** et **Sainte Apolline**, peintures enchâssées sous des architectures à décor polylobés**; Saint Guilhem** est représenté en abbé* *bénédictin (alors qu'à sa mort en 812 il n'était que simple moine); **Sainte Apolline** tient l'instrument de son martyre, une longue tenaille.*

Saint Guilhem can designate a place or a person. Is it a painting that is located in Saint Guilhem, or does it represent Saint Guilhem and Sainte Apolline?

A study of the position and the semantic class of arguments in the relationship: *subject-verb-object*, provides clues for resolving this ambiguity, the principle that the topic is the subject of the sentence, what is known as the word about the phrase, what is said of the theme.

In the above example the verb **representing** contains the feature [Repr: +], which links it with the REPRESENTATION class. In the absence of other significant indices, it can thus be inferred that the purpose of the sentence is "representation" and Saint Guilhem and Sainte Apolline do not designate places, but rather the representation.

3.3 Semi-automatic generation of a domain ontology

The knowledge gathered on an artefact is necessarily partial: it is only valid for a period of time and therefore cannot be limited to a descriptive grid designed for one specific application.

Knowledge is scalable, cultural heritage artefacts have a past, a present and perhaps a future; they undergo transformations over time.

However, we have seen above that the extraction of information in our case must correspond to precise specifications. We are thus faced with two requirements: on the one hand to populate a database defined by a specific inventory description system, on the other hand, to meet the requirements of a knowledge management system.

To satisfy the first requirement, it is essential that the information found by the extraction can be adjusted (if necessary) and validated by an expert.

To satisfy the second item, the validated information, consisting of descriptors and their relationships that describe the tangible and intangible aspects of the artefact, will have to be fed into a domain ontology, which is more extensive and extensible. This provides the necessary openness and sharing of knowledge, as defined by Gruber, *"an ontology is an explicit and formal specification of a conceptualization that is the consensus"* [13].

In the context of cultural heritage artefacts, which is the one that interests us, the description will focus on how an object was manufactured, by whom, when, for what purpose, it will focus on its transformations and travels, its conservation status and materials used for this purpose. One can see that a number of concepts are emerging such as: Time, Place, Actor (Person), state of preservation. Intuitively, one suspects that some of these concepts can be related to each other, such as conservation status and time, transformations and time, travels and place, transformations and owner.

The ontology CIDOC-CRM presents the formalism required for reporting of relationships that can be implemented in time and space. The heart of CIDOC-CRM consists of the entity expressing temporal dependence between time and various events in the life of the artefact.

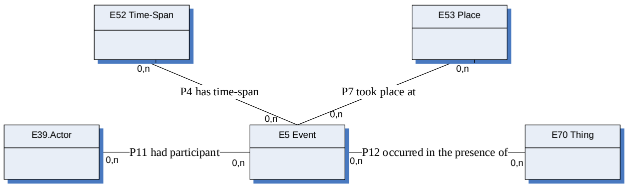


Fig. 2. Modelling event in CIDOC-CRM, from [14].

For clarity and easier reading by the user accustomed to the nomenclature of SDI we have, based on CIDOC-CRM, created a model that defines equivalences between the different fields of SDI and certain classes of CIDOC-CRM (Figure 3).

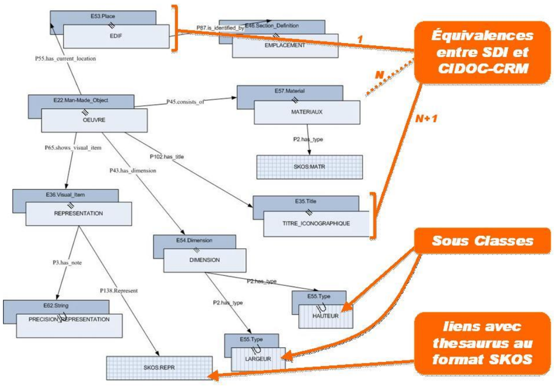


Fig. 3. The CIDOC-CRM classes and equivalence with the SDI

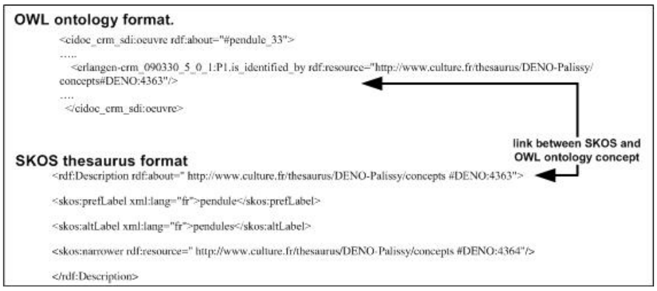
To integrate the thesaurus defined by SDI and make links between this thesaurus and CIDOC-CRM model we decided to use the SKOS [15]. SKOS format is a structured representation format for thesauri, taxonomies and, generally, for any type of controlled vocabulary. It uses the RDF formalism and thus, it defines the concepts and links between them using the properties for the identification, description, structure and organization of conceptual schemes.

An extract from the thesaurus DENOMINATION (NAME) in SKOS format, formalizes the representation of the concept DENO: 4363 described by the term pendule (clock) as preferred lexical form and terms pendules (clocks), horloge à poser as a lexical alternatives forms. It has a generic concept DENO: 4351 and a specific concept DENO: 4364.

<rdf:Description rdf:about=" http://www.culture.fr/thesaurus/DENO-Palissy/concepts #DENO:4363">

```
<skos:broader rdf:resource=" http://www.culture.fr/thesaurus/DENO-Palissy/concepts #DENO:4351"/>
<skos:prefLabel xml:lang="fr">pendule</skos:prefLabel>
<skos:altLabel xml:lang="fr">pendules</skos:altLabel>
<skos:altLabel xml:lang="fr">horloge à poser</skos:altLabel>
<skos:narrower rdf:resource=" http://www.culture.fr/thesaurus/DENO-Palissy/concepts #DENO:4364"/>
</rdf:Description>
```

The link between concept represented in OWL format and concept from thesaurus in SKOS format is established in the case of concept DENOMINATION, by the relation P1.is_identified_by defined in CIDOC-CRM; this link is shown in Figure 4.



OWL ontology format.

```
<cidoc_crm_sdi:oeuvre rdf:about="#pendule_33">
.....
  <erlangen-crm_090330_5_0_1:P1.is_identified_by rdf:resource="http://www.culture.fr/thesaurus/DENO-Palissy/
concepts#DENO:4363"/>
....
 </cidoc_crm_sdi:oeuvre>
```

SKOS thesaurus format

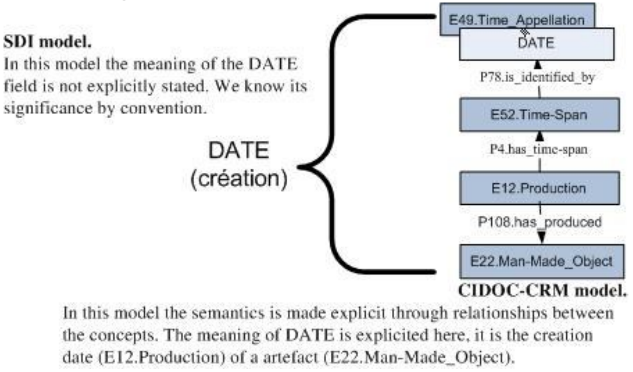
```
<rdf:Description rdf:about=" http://www.culture.fr/thesaurus/DENO-Palissy/concepts #DENO:4363">
<skos:prefLabel xml:lang="fr">pendule</skos:prefLabel>
<skos:altLabel xml:lang="fr">pendules</skos:altLabel>
<skos:narrower rdf:resource=" http://www.culture.fr/thesaurus/DENO-Palissy/concepts #DENO:4364"/>
</rdf:Description>
```

Figure 4. Link between CIDOC-CRM formalised in OWL and the thesaurus in SKOS format.

The transition from the model defined by the inventory descriptive system to the CIDOC-CRM ontology (cf stage 6, figure 1), will be done by searching through the correspondence between the fields of inventory descriptive system, whose contents can be regarded as an instance of one of the classes of the CRM ontology.

For cases where this correspondence can not be made, because the information does not exist in the inventory description system, it will have to be retrieved from the transcribed text, provided that the speaker has record such kind of information. Otherwise it will have to be input when the information extracted automatically by the system is validated.

Correspondence between the implicit information in the SDI and the explicit information of CIDOC-CRM model is shown in Figure 5.



SDI model.
In this model the meaning of the DATE field is not explicitly stated. We know its significance by convention.

CIDOC-CRM model.
In this model the semantics is made explicit through relationships between the concepts. The meaning of DATE is explicited here, it is the creation date (E12.Production) of a artefact (E22.Man-Made_Object).

Figure 5. Example of correspondence between SDI and CIDOC-CRM

As shown in this figure, to make a transition of a SDI information to CIDOC-CRM, the system must be able to associate one or more concepts that can be an event (creation), tangible or intangible object, a date or period, which are connected through specific relations.

4 Experiments and results

The application that we propose is still in prototype stage; it is therefore too early to provide a real experience feedback, which would require the operation of our system.

Thus, we present the experiments we have conducted so far with the prototype version of our system and with the help of three researchers familiar with cultural heritage as well as the area of inventory and SDI. Two of the three researchers are female, one of which has a regional accent, while the other speaks with no accent. The third, male researcher speaks with accent.

The dictations were performed in real conditions in a noisy environment. We asked each researcher to verbally describe three objects.

The oral descriptions were transcribed into text. The results are quite satisfactory; the concordance between the original content and the content in the automatically transcribed texts varies between 90 and 98%.

Before presenting them to the module for the extraction of information, the transcribed texts have been corrected by the researchers. For each result of extraction of information, we measured Precision, Recall and F-score, which are presented in the table below.

In order to clarify the presentation we have assigned a letter to designate each speaker: A for the woman speaking with an accent, B for the woman with no accent and the letter C for the man.

TABLE 1. RESULTS OF EXTRACTION OF INFORMATION

Researcher	Precision	Recall	F-score
A	0,898	1	0,943
B	0,854	0,946	0,897
C	0,903	0,94	0,921

Our experiments are not numerous enough to supply a more reliable statistical study, nevertheless the obtained results are sufficiently promising to encourage us to continue developments of our system.

For the moment our system is elaborate for the French language.

Below is an example of the description of a painting performed by a researcher of cultural heritage. The first text is the result from the voice recording transcript. You can see the errors marked in bold.

Et le Damiani église Saint-Sauveur. Tableau représentant saint Benoît d'Aniane et saint Benoît de Nursie offrant à

Dieu le Père la nouvelle église abbatiale d'Aniane. Ce tableau est situé dans le coeur et placé à 3,50 m du sol. C'est une peinture à l'huile sur toile encadrée et 24 en bois Doré. Ça auteure et de 420 cm sa largeur de 250 cm. Est un tableau du XVIIe siècle. Il est signé en bas à droite droite de Antoine Ranc. Est un tableau en mauvais état de conservation un réseau de craquelures s'étend sur l'ensemble de la couche picturale.

The second is the text after correction. You can consult the translation of this text in English in appendix.

The results outcomes from module of the Extraction of Information are marked in bold.

Ville d’COM{Aniane} EDIF{église Saint-Sauveur}. PREPR{DENO{Tableau} représentant REPR{ saint Benoît d'Aniane] et REPR {saint Benoît de Nursie} offrant à REPR{Dieu le Père} la nouvelle église abbatiale d'Aniane}. Ce tableau est situé EMPL{ dans le choeur et placé à 3,50 m du sol}. C'est une peinture à MATR{ l'huile sur toile} encadrée d’un cadre en MATR{bois doré}. Sa DIMS{hauteur est de 420} cm sa DIMS{ largeur de 250 cm}. Est un tableau du SCLE{XVIIe siècle}. Il est signé en bas à droite de AUTR{Antoine Ranc}. PETAT{ Est un tableau en mauvais état de conservation un réseau de craquelures s'étend sur l'ensemble de la couche picturale}.

Where:

COM = Commune, EDIF = Edifice, REPR =Representation, PREPR = Precision on the representation, EMPL = Place, MATR = Materials, DIMS = Dimension, SCLE = Century, AUTR = Author, PETAT = Precision on the state of preservation.

The linguistic analysis, information extraction and ontology creation are done using the second file, as shown schematically in Figure 6.

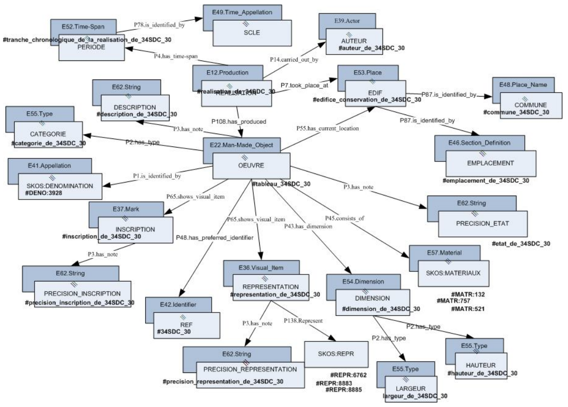


Fig. 6. Example of ontology of a work after a description dictation.

5 Conclusion

The originality of our voice recording system developed to support the acquisition of knowledge of cultural heritage is the link between two areas of research, which were until now developing parallel to each other: signal processing and automatic language processing. Our experiments have been successful and confirm the technical feasibility and usefulness of such applications.

Modelling of knowledge as an ontology and ontological cooperation will provide flexibility and scalability to our system, e.g. extending the scope of the CIDOC-CRM model to model the spatio-temporal knowledge, by adding geospatial information such as topology, directions, distances, location of an artefact relative to reference locations. The recent work of LIG (Laboratoire d'Informatique de Grenoble) and in particular the model ONTOAST [16] seem very interesting in this regard. In the context of ontological cooperation arises the problem of coherence among distributed ontologies, who we believe can be resolved by means of the cognitive agents.

In the future, it might be useful to incorporate a speech acquisition control mechanism, in the form of a man-machine dialogue. Thus the speaker would have a real-time feedback on the machine’s understanding. This implies in our case the possibility to implement the transcription and information extraction system on a mobile platform.

The OWL format for the creation of the ontology we use ensures its compatibility with the standards of the semantic web. It allows for an easy integration with inference and inquiry systems, thereby facilitating its future use in both scientific and community applications, such as search engines, artefact comparison platforms or the exchange of knowledge with other ontological structures.

References

[1] du Château S, (2010), SIMPLICIUS, Système d’aide au management des connaissances pour le patrimoine culturel, Thèse en Informatique, Université Lyon3.

[2] Iacovella A., Bénel A. et al., (2003) Du partage de corpus documentaires structurés à la confrontation de points de vue, Dossier d’identification d’une équipe projet CNRS STIC, Juillet 2003.

[3] Amidon D. M. (1997) Innovation Strategy for The Knowledge Economy, Butterworth Heinemann.

[4] Mercier-Laurent E. (2007), Rôle de l'ordinateur dans le
processus global de l'innovation à partir de connaissances, Mémoire
d’Habilitation à Diriger la Recherche en Informatique, Université Lyon3.

[5] Ibekwe-SanJuan F. (2007), Fouille de textes : méthodes, outils et applications, Hermès-Lavoisier, Paris-London, 2007, 352p.

[6] Grishman, R. (1997). Information Extraction: techniques and challenges. Information Extraction (MT Pazienza ed.), Springer Verlag (Lecture Notes in computer Science), Heidelberg, Germany.

[7] Aït-Mokhtar S., Chanod J.-P. et Roux Cl. (2002). Robustness beyond shallowness : incremental deep parsing. Natural Language Engineering, vol. (8/2-3) : 121-144.

[8] Verdier H. (1999), Système descriptif des objets mobiliers. Paris, 1999.- Editions du Patrimoine.

[9] Hagège C., Roux C. (2003), Entre syntaxe et sémantique: Normalisation de la sortie de l’analyse syntaxique en vue de l’amélioration de l’extraction d’information à partir de textes, TALN 2003, Batz-sur-Mer, 11–14 juin 2003.

[10] Caroline Brun, Caroline Hagege Semantically-Driven Extraction of Relations between Named Entities CICLing 2009 (International Conference on Intelligent Text Processing and Computational Linguistics), Mexico City, Mexico, March 1-7, 2009

[11] Enkvist N.E, (1976), Notes on valency, semantic scope, and thematic perspective as parameters of adverbial placement in English". In: Enkvist, Nils E./Kohonen, Viljo (eds.) (1976): Reports on Text Linguistics: Approaches to Word Order.

[12] Ho-Dac L. (2007), La position Initiale dans l'organisation du discours : une exploration en corpus. Thèse de doctorat, Université Toulouse le Mirail.

[13] GRUBER T. R. (1993). A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5, 199.220.

[14] Crofts N. (2007), La norme récente ISO 21127: une ontologie de référence pour l'échange d'infomations de patrimoine culturel, Systèmes d'informations et synergies entre musées, archives, bibliothèques universités, radios et télévisions, Lausanne – 2007.

[15] Miles A et Brickley D., (2005). SKOS Core Guide W3C Working Draft 2. novembre 2005. http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/

[16] Miron A., Gensel J., Villanova-Oliver M., Martin H.,"Relations spatiales qualitatives en ONTOAST pour le Web semantique geospatial", Colloque International de Geomatique et d'Analyse Spatiale (SAGEO2007), Clermont-Ferrand, France,18-20 June 2007.

APPENDIX

The translation of the french description :

City Aniane church Saint-Sauveur. Painting representing Saint *Benoît d'Aniane* and Saint Benoît of Nursie offering to God the Father the new abbey church of Aniane. This Painting is situated in the choir and placed in 3,50 m above the ground. It is an oil painting on canvas framed in a gilt wood frame. His height is 420 cm width is 250 cm. It is a painting of the XVIIth century. He is signed bottom on the right by Antoine Ranc. It is a picture in a poor state of preservation a cracks network spread throughout the entire painting area.

GAMELAN: A Knowledge Management Approach for Digital Audio Production Workflows

Karim Barkati [1] and **Alain Bonardi** [2] and **Antoine Vincent** [3] and **Francis Rousseaux** [4]

Abstract. The ongoing french research project GAMELAN aims at demonstrating how knowledge management principles and technics could serve digital audio production workflows management, analysis and preservation. In this position paper, we present both the production stakes of such an approach and the technical and scientific knowledge strategies we are developing, through the coupling of both knowledge and process engineerings.

1 INTRODUCTION

GAMELAN[5] is an ongoing french research project, which name means "An environment for management and archival of digital music and audio". It aims at answering specific needs regarding digital audio production – like interoperability, reusability, preservation and digital rights management – while striving to settle knowledge management issues, combining knowledge engineering with process engineering.

1.1 Digital audio production stakes

From a social standpoint, the large and growing number of users of audio environments for personal or applied production makes this field one of the richest in evolution. However, the complexity of the production management is a well known effect in the community and is often described as an inconsistency between the tools used. Indeed, the industry provides more and more powerful tools but regardless of global usage: users combine multiple tools simultaneously or constantly alternating from one tool to another.

From a legal standpoint, there is a real problem of contents tracking, given their multiple uses or changes in production. Till now, audio production systems keep no operational track that would allow following up the rights associated with each element.

Thereby, the GAMELAN project goals spread on four levels:

Production Environments — Keep track of all actions, since the starting material to finished product; organize production elements (files, software) in structures included as components of the environment; formalize the knowledge generated during the process.

Preservation Strategies — Use the production environment as a platform for preservation; extract structures and knowledge to simplify future access to the environment; apply OAIS[6] methodologies, allowing reuse of the environment and its components.

Reuse of productions — Restructure the production elements with new objectives, adding other materials and editing the links and the overall structure (*repurposing*, back-catalog rework); use subsets of the environment to generate new environments and facilitate the process deconstruction and reconstruction for intentions analysis.

Digital rights management — Enable traceability of content on a production to manage user rights. Detect and warn for missing DRM information (artists name, location, person in charge of the production) during the production process itself, from *creation patterns* specification.

1.2 Use cases and technical functionalities

The technical goal of GAMELAN research project is to create a software environment (also called GAMELAN), integrating musical and sound production softwares, and able to fully describe the workflow, from source to final product. GAMELAN conforms to Open Source Initiative criteria, is free, and defines guidelines to allow the meta-environment to extract specific software information.

We elaborated three use cases relying on the different expertises of project partners.

CD production at EMI Music — On digital audio workstations ("DAWs", like ProTools or Audacity). CD production workflow involves recording, mixing and mastering steps. The main related test case consists in removing a particular track from a song for *repurposing* (like the voice for karaoke edit), the second one in identifying all contributors name to ease rights management.

Acousmatic music creation at INA/GRM — Also on DAWs. Workflow involves at least mixing and editing steps. The main test case consists in identifying which file is the "final mix" for archiving, the second one in identifying eventual versions varying only in format (compression, number of channels, etc.).

Patch programming at IRCAM — On audio programming environments like Max/MSP, for real-time interactive works. The main test case consists in visualizing structural changes of patches (subpatches and abstractions calls) for genetic analysis, the second one in indexing control parameters values for centralized fine-tuning.

As a meta-environment, GAMELAN traces data during the production process and utilizes formalized knowledge upon collected

[1] Institut de Recherche et Coordination Acoustique/Musique (IRCAM), France; email: karim.barkati@ircam.fr

[2] Paris 8 University, EA 1572 – CICM & IRCAM, France; email: alain.bonardi@ircam.fr

[3] Université de Technologie de Compiègne, Laboratoire Heudiasyc, UMR 7253 CNRS, France; email: antoine.vincent@hds.utc.fr

[4] Reims Champagne-Ardenne University, CReSTIC EA 3804 & IRCAM, France; email: francis.rousseaux@univ-reims.fr

[5] http://www.gamelan-projet.fr, "*Un environnement pour la Gestion et l'Archivage de la Musique Et de L'Audio Numériques*".

[6] Open Archive Information System.

data, both during and after production time. Main technical functionalities are tracing, acquisition, ingestion, reasoning, requesting, browsing, file genealogy visualisation, integrity and authority checking, and archiving.

1.3 Knowledge management issues

On the way to GAMELAN's goals, we identified three main aspects:

Archival issue — How to work out a representation of the musical objects that allows to exchange, take back or reproduce them, while each musical environment is most of the time particular and contingent? Furthermore, this representation has to be abstract in that it must be general enough to be valid and usable in other contexts, and concrete enough to contain the information necessary for the reuse of objects.

Cognitive issue — How to characterize and explain the part of the knowledge necessary to understand and reuse tools and their settings, while the creation process mobilizes a set of intentions and knowledge rom the author that are only implicitly transcribed in its use and settings? This information will be included in the abstract representation elaborated in the previous archival issue.

Technical issue — How to make explicit the knowledge about the creative process as well as production tools despite the heterogeneity of abstraction levels of objects and the fragmentation of objects and tools that do not communicate with each other? Moreover, the worked out representations and models should to be exploitable in a technical environment that offers the user the ability to interact with data and structures.

So, the GAMELAN project addresses issues relevant to several knowledge fields:

- Digital archiving through intelligent preservation;
- Management and capitalization of knowledge;
- Process engineering and knowledge engineering.

The goal here is to define a trace engineering, that is to say intermediate objects built by the composer or producer and associated settings. To reach this goal, we divided the global work into three main tasks presented in the paper:

- Production process tracking, that should be agnostic;
- Professional knowledge models, that may be hierarchical;
- Work and process representation, that should allow both visualizing, querying and editing.

2 PRODUCTION PROCESS TRACKING

The first step we consider deals with gathering data, through logging user interaction events and collecting contextual information.

2.1 The GAMELAN meta-environment

The applicative objective of the project is to create a meta-environment for music and audio production, capable of integrating any type of production software, and able to fully describe the workflow, from source to final product.

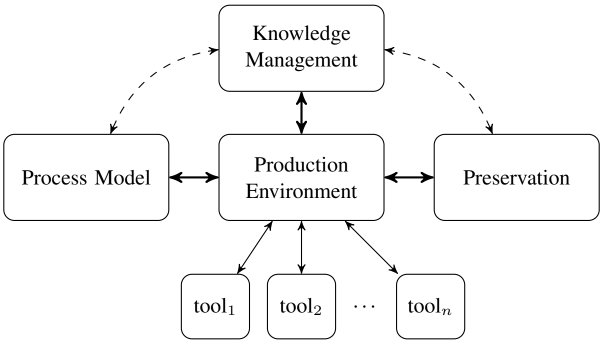


Figure 1. High-level technical architecture of GAMELAN.

2.1.1 High-level technical architecture

The technical architecture relies on the *production environment*, which includes various digital audio production tools at work in the process, as shown on Fig. 1.

It is based on predefined *process models* to measure and qualify the steps and operations performed during a particular process, related to a unit of *knowledge management* that provides methods for evaluating this process and provide at any time the user an evaluation of the current process and context sensitive help. Therefore, it aims at providing at all times an overview of the entire process in terms of progress, quality, and outcome.

Users should be able to control the interaction of this feedback with their own work, which implies non-intrusiveness and transparency for the meta-environment.

Finally, an *archive unit* will allow an smart preservation of digital objects, keeping the "footprint" of the entire process to allow full traceability. This unit will be based on the OAIS MustiCASPAR server developed within the CASPAR project [8], and adapted to the preservation of the production process.

2.1.2 Operational tracking

Considering the dynamic nature of knowledge is a key issue in knowledge engineering. Indeed, whatever the quality of the modeling process and the quantity of knowledge collection, resulting knowledge traces are then to be mobilized in contexts never completely predictable and these inscriptions will report a reality that has evolved by itself. This is the reason why we designed an operational tracking process as agnostic as possible, through messaging, tracing and logging.

The messaging part relies on an open-source standard commonly used in the computer music community, namely OSC[7], developed at UC Berkeley [15], which is a communication protocol for modern networking technology, with a client/server architecture (UDP and TCP).

In order to produce usage data [10, 13], we hacked open-source domain production softwares, like Audacity[8], to send a complete OSC message each time the user performs an action through the software, build with:

- Application name

[7] http://opensoundcontrol.org/, *Open-sound control.*

[8] http://audacity.sourceforge.net, an open source software for recording and editing sounds.

- Application version number
- Time stamp
- Function name
- Function parameters

We developed a tracing and logging application (the tracker) that both logs every message received during the production, only adding a reception time stamp, and keeps track of every version of modified files, tracking also file system messages, for file genealogy analysis and preservation purposes.

Hereafter are excerpts of log files, reduced to fit here (timestamps and/or other information are removed).

OSCMessages.txt

```
audacity 1.3 FileNew
audacity 1.3 FileSaveAs  test.aup
audacity 1.3 ImportAudio test.aup noise.wav
audacity 1.3 ImportAudio test.aup clicks.wav
audacity 1.3 Selection mix.aif ”noise”, ”clicks”; Begin=”1.9314”; End=”10.0114”
audacity 1.3 ExportAudio test.aup mix.aif
audacity 1.3 FileClosed test.aup
```

CurrentApplication.txt

```
2012-07-09 10:09:36544633 +02 ApplicationActivated net.sourceforge.audacity
2012-07-09 10:09:36582045 +02 ApplicationActivated com.apple.dt.Xcode
2012-07-09 10:09:36593654 +02 ApplicationActivated com.apple.finder
```

FolderState.txt

```
folder-state 0        2012-07-10 16:22:58961547 +02
2012-01-20 18:07:65253000 +01        noise.wav
2012-01-20 18:07:65253000 +01        clicks.wav
...
folder-state 7        2012-07-10 16:23:58981517 +02
2012-01-20 18:07:65253000 +01        noise.wav
2012-01-20 18:07:65253000 +01        clicks.wav
2012-07-10 16:23:58980000 +02        test.aup
...
folder-state 21        2012-07-10 16:23:59005107 +02
2012-01-20 18:07:65253000 +01        noise.wav
2012-01-20 18:07:65253000 +01        clicks.wav
2012-07-10 16:23:59005000 +02        mix.aif
2012-07-10 16:23:58980000 +02        test.aup
```

2.1.3 *Manual informing*

Knowledge management as defined in our project requires further information that can not be inferred from the software activity logging. Indeed, a set of primary contextual information must be given by a human operator, like the user's name and the title of the work being produced. But a design dilemma immediately appears: on the one hand, the more contextual information feeds the system, the more informative might be the knowledge management, but on the other hand, the more a system asks a user to enter data, the more the user may reject the system [4].

In our case, the balance between quantity and quality of information has to be adjusted in a close relationship with the strongly-commited ontology we are incrementally developing with domain experts [14] and presented thereafter.

Temporal modalities have also to be anticipated in the information system, since the operational manual informing phase can be entered either at the same time that the production phase or temporally uncoupled, either by the producing user (e.g. a composer) or by an external agent (e.g. a secretary). Moreover, crucial missing data detection by the knowledge management system is a key feature of the project, as information integrity checking.

2.2 Managing knowledge flows

2.2.1 *Ontology-driven KM*

As we saw, the manual informing part of the system strongly depends of the domain ontology, but this is not the only part. Indeed, knowledge management depends on the ability to transform data and information into knowledge, according to Ackoff's model [1], and it turns out that ontologies are key tools in this transition process [9, 6]. We incrementally developed a strongly-commited differential ontology dedicated to audio production, dipping in productions with experts, in the OWL formalism.

In our system, except for the operational tracking that has to remain agnostic, the ontology drives all functional modules:

Data — The informer module we saw previously for contextual user data, especially for the entry interface design;

Information — The preprocessing module that prepares raw data (both usage data and user data) according to the ontology;

Knowledge — The semantic engine reasoning on the preprocessed information, and allowing requests;

Understanding — The browser module for data browsing and edition, and the viewer module that provides global graphical representations, like file genealogy trees.

The central position of the ontology comes from its semantic capabilities and justifies deep research toward professional knowledge modeling in music.

2.2.2 *Production process tracing*

We distinguish between *user data* and *usage data*. The former corresponds to the manual informing data and the latter to the automatic tracking data. In the computer music field, this production process tracing has never been done yet. We asked for use cases to domain professionals (cf. Section 1.2) in order to reproduce relevant user interaction with the production meta-environment.

This strategy aims several beneficiaries and time horizons:

In the immediate time of production — The composer, audio producer, may turn back its own work during the production, to explore various options or correct the undesirable consequences; it can be for example a selective "undo" instruction given to cancel an operation; it is also, for the composer or the sound engineer an opportunity to see and understand the overall work of composition or production.

In the intermediate time of collection — The composer, or the institution that manages its works, may return on a given work to recreate or reuse the content components.

In the long term preservation — The work becomes a memory and a relic, the challenge is to preserve the artistic and technical information to understand, interpret and re-perform.

3 PROFESSIONAL KNOWLEDGE MODEL

3.1 Modeling context

It is common to begin the modeling phase by a corpus analysis, usually from a collection of candidates-documents selected depending on their relevance [12]. But in our case study, we have no written document that can provide support to terms selection: vocabulary, and by extension, all production work relies on musical practices that are acquired more by experience than by teaching. Indeed, every musical work is a *prototype* in the sense of Elie During, as "the most perfect example, the more accurate", where each creation is an object "ideal and experimental": this uniqueness leads to a possibly infinite number of ways to create [5]. Thus, to achieve this essential phase of study, we design ourselves our corpus, which is rather unusual, by following several musical productions to find out invariants.

We do not seek to explain sound nor music (the *what*, like MusicXML kind of languages) but the way it is produced (the *how*), *i.e.* a formal language for audio production process. This language is devoted to the representation of what we might call the "musical level", referring to the "knowledge level" of Allen Newell: we want to represent the work at the right abstraction level, neither too concrete because too technology dependent and therefore highly subject to obsolescence, nor not enough because information would be too vague to be usable [11].

3.2 Main test cases

The GAMELAN project embraces various creative practices, related to the partners core business who defined three main test cases:

IRCAM *Recovery assistance and synthesis of information from one phase to another of a record.* We followed the recording and editing situation of the piece "Nuages gris" of Franz Liszt in the "Liszt as a Traveler" CD played by pianist Emmanuelle Swiercz. — Identify and represent the work of the sessions in two dimensions by time and by agent, all the events of one session (creation, update, export), and the dependencies of import and export files between sessions.

INA/GRM *Identification of files that have contributed to the final version of a work.* We log every DAW operation of composer Yann Geslin during the composition of a jingle. — Ensure that the file called "Final-Mixdown" is actually the one that produced the last audio files of the work; identify possible format changes (stereo, 8-channel, mp3); identify the intermediate versions.

EMI Music *Recovery and edit of past productions.* We plan to test the replacement of the drum from a recording made under GAMELAN. — Accurately identify which tracks to replay; substitute to an identified track for another; replay the final mix session with the replaced tracks.

3.3 Production process modeling

To create the representation language of the production process, we apply the *Archonte*[9] method of Bachimont [2].

Our production process modeling work followed three steps:

1. Normalization of the meanings of selected terms and classification in an ontological tree, specifying the relations of similarity between each concept and its father concept and/or brothers concepts: we then have a *differential* ontology;
2. Formalization of knowledge, adding properties to concepts or constraining relation fields, to obtain an *referential* ontology;
3. Operationalization in the representation language, in the form of a *computational* ontology.

After a phase of collection of our corpus and the selection of candidate terms, we took the first step in the form of a taxonomy of concepts, in which we strived to maintain a strong semantic commitment in supporting the principles of the differential semantics theory presented thereafter. This taxonomy has been performed iteratively, since it is dependent on our participation in various productions. Thus, at each new integration to the creation or the updating of a work, we flatten and question our taxonomy and term normalization, in order to verify that the semantic commitment is respected.

For incremental development and testing, we divided the ontology in two parts: the one as a model, with classes and properties, uploaded on a dedicated server `icm.ircam.fr`[10], the other as conform data sets, with individuals, uploaded on an OWL server (OpenRDF Sesame plus the OWLIM-Lite plugin) `gsemantic.ircam.fr`[11]. For common features, we import standard ontologies, like vCard[12] for standard identity information, so we will only detail the making of the domain ontology in this article.

3.4 The differential approach

In short, the differential approach for ontology elaboration systematically investigates the similarity and difference relations between each concept, its parent concept and its sibling concepts. So, in developing this structure, we tried to respect a strong ontological commitment by applying a *semantic normalization*, that is to say that for each concept, we ask the four differential questions of Table 1.

S. w/ P.	– Why does this concept inherit from its parent concept?
D. w/ P.	– Why is this concept different from its parent concept?
S. w/ S.	– Why is this concept similar to its sibling concepts?
D. w/ S.	– Why is this concept different from its sibling concepts?

Table 1. The four differential questions.

To realize practically this semantic normalization task, we used softwares DOE[13] [3] and Protégé[14], for both concepts and relations taxonomies building, refining and exporting (RDFS, OWL, etc.). At the end of this recursive process, we obtained a domain-specific differential ontology (see excerpt on Fig. 2), where the meaning of all terms have been normalized and that allows to develop the vocabulary needed for the next steps to reach the development of the representation language of the audio production process.

As a result of the differential method, domain vocabulary mostly lies in leaves of such an ontological tree: *work, performance, version; connection, graphical object, track, region; association, enterprise, institute; musical score, instrument, brass, strings, percussions, winds; effect box, synthesizer; file, session file, program; create, delete, edit; content import, content export; listen, play, work session, current selection*; etc. A large set of properties completes this domain ontology.

4 WORK AND PROCESS REPRESENTATIONS

The idea of such a meta-environment as GAMELAN, viewed as a trace-based system (cf. Fig. 3) meets clear needs in the community, as mentioned before. Moreover, while the operational meta-environment is still under development, our ontological work already points to the solution of various scientific challenges:

- Representation language for managing the creation process;
- Description language for representing the content of a work, in the diversity of its components;
- Integration of both languages in a single control environment.

[9] ARCHitecture for ONTological Elaborating.

[10] http://icm.ircam.fr/gamelan/ontology/2012/07/10/SoundProduction.owl

[11] http://gsemantic.ircam.fr:8080/openrdf-workbench/repositories/

[12] http://www.w3.org/Submission/vcard-rdf

[13] http://www.eurecom.fr/~troncy/DOE/, *Differential Ontology Editor*.

[14] http://protege.stanford.edu/

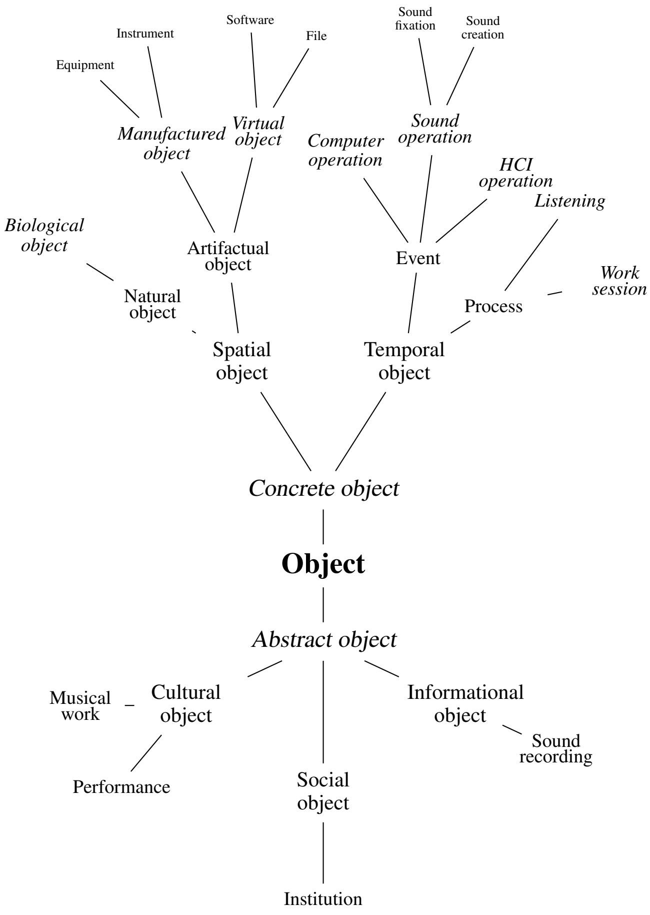


Figure 2. Excerpt from the differential taxonomy

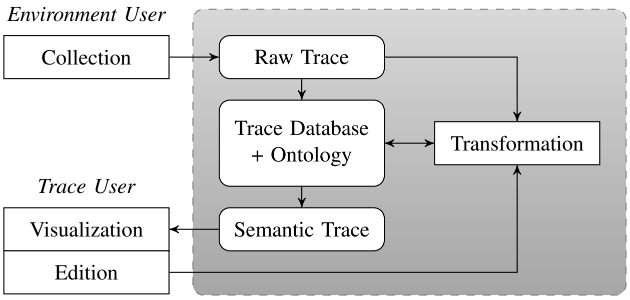


Figure 3. Schema of GAMELAN trace-based system.

4.1 Content description language

The descriptive approach is not to keep the content stored, because content is usually partial, incomplete or poorly defined (*ad hoc* formats, imperfect knowledge of it, etc.). Rather, it is better to retain a description of the content that enables to reproduce it. The description may include the main points to reproduce, the author's intention to comply [7], the graphical appearance, etc.

So, the description of the content of a work is an approach increasingly adopted in response to the technical complexity (mostly digital) of content: instead of maintaining a technical object that we may no longer know how to handle, we shall construct a description to reinvent this object with the tools we will have at hand when the time comes. Such a description necessarily introduces a deviation from the original: the challenge being that this difference does not affect the integrity nor the authenticity of the work.

The main question is how to determine such a description language. The score used in the so-called classical music, is a good example of such a language. Instead of stepping on the impossible task to keep a musical object before recording techniques, musicians preferred to keep *the instructions to create it*. Now, the complexity of the works, the mutability and fragility of digital imply that it is impossible to guarantee that a technical object will still be executable and manipulated in the future.

Several approaches are possible, but some semiotic and logic work has to be conducted to identify such a description stage:

- Semiotic, because it is necessary to characterize the objects mobilized in a production, define their significance and propose an associated representation;
- Logical, since this representation must be enrolled in a language for control actions in the proposed meta-environment.

4.2 Time axis reconstruction

The proposed description must also be temporal and allow browsing of the different states of the work. The representation of time must be done in terms of versions, traces of transformations, to offer the user the ability to revert to previous states and build new states by reusing some earlier versions of objects composing his work.

Indeed, in the final stage of production, archiving of music and sound production is generally confined to the archiving of a final version ("mastered"). Whereafter it is clearly impossible from this single object to trace the production history, nor to take back and modify the process in a different perspective, while informed musical remix is a clear identified need with *repurposing* aims.

This lead us to ensure strong timing properties through our trace-based system, not only time stamping user events from the production tools when emitting messages, but also independently time stamping a second time these events in the logging module when receiving messages. This allows us to reconstruct the time axis of the production safely.

4.3 Database browsing and timeline visualization

Here, the digital archival issue of provenance should be avoided or at least diminished upstream the ingest step, thanks to knowledge management. The GAMELAN meta-environment is intended to be able to detect crucial missing information by reasoning on the combination of software traces and user information. This important features, dedicated to the trace user, are carried out through common knowledge management tools, namely:

- Domain ontology;
- Trace database;
- Query engine;
- Semantic repository;

Besides, a timeline visualization tool brings a global view to help the answers understanding, typically showing the genealogy of the files used during the production. For example, GAMELAN can infer which files were used to compose a mixed file, hierarchically, and also deduce which is the "last mix" in a set of file; this kind of knowledge is of prime importance when a composer or a producer decides to remix a work years latter.

4.4 Creation patterns

Now that our ontology has reached a decent level and stabilizes, we enter a second phase of our ontological research: *creation patterns* design. These patterns will define audio creation acts. The use of these patterns will allow to represent a set of actions with a musical meaning, incorporating the vocabulary developed in the ontology.

Technically, we chose to stick to the OWL formalism, instead of switching to process languages like BPMN, to describe and analyse some relations between ontology objects with domain experts, especially for relations that they stressed as being of prime importance regarding test cases.

From these creation patterns, we intend to derive query patterns, in an automated way as much as possible. Indeed, a common formalism between the ontology and the creation patterns ease both reuse of the vocabulary during the pattern design phase and the translation into query patterns, especially when using compliant query languages like SPARQL as we do on our Sesame repository.

Knowledge will be then bilocalized: on the semantic repository side for objects of the trace database, and on the integrity and authenticity checker side for the formalized relations of the query patterns base.

5 CONCLUSION

Along this position paper of the GAMELAN research project, we presented how a knowledge management approach for digital audio production workflows could be of great utility at several time horizons: in the immediate time of production, in the intermediate time of collection, and in the long term preservation.

We also detailled how we are combining a trace-based architecture and an ontology-driven knowledge management system, the latter being build upon differential semantics theory. Technically, semi-automatic production process tracking feeds a semantic engine driven by production process ontology levels. Clearly, this requires both knowledge engineering and process engineering but also digital preservation methods awareness.

At last, the project will provide the following results:

- A software environment, published as free software, used to drive selected production tools and capable to accommodate to other tools later on, thanks to its openness;
- A representation and description language of manipulated content, including their temporal variation and transformation;
- A representation language of the digital audio creation process.

ACKNOWLEDGEMENTS

The GAMELAN project is funded by French National Research Agency. It started in 2009 and will end in 2013. The partners are:

- IRCAM (*Institut de Recherche et Coordination Acoustique/ Musique*),
- Heudiasyc – UMR 7253 CNRS UTC,
- INA (*Institut National de l'Audiovisuel*),
- EMI Music France.

REFERENCES

[1] R.L. Ackoff, 'From data to wisdom', *Journal of applied systems analysis*, **16**(1), 3–9, (1989).

[2] B. Bachimont, 'Ingénierie des connaissances', *Hermes Lavoisier, Paris*, (2007).

[3] B. Bachimont, A. Isaac, and R. Troncy, 'Semantic commitment for designing ontologies: a proposal', *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, 211–258, (2002).

[4] H. Barki and J. Hartwick, 'Measuring user participation, user involvement, and user attitude', *Mis Quarterly*, 59–82, (1994).

[5] Elie During, 'Entretien avec Franck Madlener', in *L'Étincelle*, ed., Ircam, Paris, (2010).

[6] D. Fensel, F. Van Harmelen, M. Klein, H. Akkermans, J. Broekstra, C. Fluit, J. van der Meer, H.P. Schnurr, R. Studer, J. Hughes, et al., 'On-to-knowledge: Ontology-based tools for knowledge management', in *Proceedings of the eBusiness and eWork*, pp. 18–20, (2000).

[7] L. Gaillard, J. Nanard, B. Bachimont, and L. Chamming's, 'Intentions based authoring process from audiovisual resources', in *Proceedings of the 2007 international workshop on Semantically aware document processing and indexing*, pp. 21–30. ACM, (2007).

[8] D. Giaretta, 'The caspar approach to digital preservation', *International Journal of Digital Curation*, **2**(1), (2008).

[9] T.R. Gruber et al., 'Toward principles for the design of ontologies used for knowledge sharing', *International journal of human computer studies*, **43**(5), 907–928, (1995).

[10] I. McLeod, H. Evans, P. Gray, and R. Mancy, 'Instrumenting bytecode for the production of usage data', *Computer-aided design of user interfaces IV*, 185–195, (2005).

[11] A. Newell, 'The knowledge level', *Artificial intelligence*, **18**(1), (1982).

[12] F. Rousseaux and A. Bonardi, 'Parcourir et constituer nos collections numériques', in *CIDE Proceedings*, pp. 133–142, Nancy (France), (2007).

[13] S. Smith, E.D. Schank, and B.M. Tyler. Instrumented application for transaction tracing, March 29 2005. US Patent App. 11/092,428.

[14] A. Vincent, B. Bachimont, and A. Bonardi, 'Modéliser les processus de création de la musique avec dispositif numérique : représenter pour rejouer et préserver les œuvres contemporaines', in *Journées francophones d'ingénierie des connaissances (accepted)*, (2012).

[15] M. Wright, A. Freed, and A. Momeni, 'Opensound control: state of the art 2003', in *Proceedings of the 2003 conference on New interfaces for musical expression*, pp. 153–160. National University of Singapore, (2003).

Knowledge Management applied to Electronic Public Procurement

Helena Lindskog
Helena.lindskog@liu.se
Department of Management and Engineering
Linköping University
Sweden

Danielle Boulanger
db@ univ-lyon3.fr
&
Eunika Mercier-Laurent
e.mercier-laurent@univ-lyon3.fr
MODEME, IAE Research Center
University Jean Moulin, Lyon3
France

Abstract

Public procurement is a knowledge-based process**.** It involves, amongst others the knowledge of needs and trends, knowledge of concerned products or services, on their evolution in time and knowledge about actors able to offer them. The knowledge of political and legal context should be also considered as well as the environmental and social impact. Electronic procurement aims in reducing the amount of paper, but also in quicker and more knowledgeable processing of proposals and decision taking. We consider procurement activity as a part of a global organizational knowledge flow. This work goal is to analyze the whole process, identify the elements of knowledge necessary for successful purchase processing and to study the contribution of AI approaches and techniques to support the above elements.

1. Public Procurement

Purchasing is one of the most important activities for any kind of organisation. Axelsson and Håkansson (1984) define three different roles for purchasing:

- The rationalization role – to buy at very competitive prices which will put pressure on supplier efficiency.
- The developing role – to monitor the technical development (product and process) in different supplier segments and to encourage the suppliers to undertake technical development projects.
- The structuring role – to develop and maintain a supplier structure with a high potential for both development and efficiency.

. While purchasing is in many ways similar for both public and private sectors, public procurement is in almost all situations and countries regulated by a specific legislation, which is stricter than the one that regulates the private sector's purchasing activities. For example, in public procurement environment the Request for Proposal (RfP) after it has been published cannot be changed, there is a possibility to appeal for suppliers if they consider themselves of being unfairly treated in the tendering process and must be taken into consideration not only economical but also political goals such as environmental and societal. The closest similarity between public procurement and private purchasing is probably in the case of acquisitions of large investments, often called project purchasing (see Ahlström, 2000; Bonnacorsi et al., 1996; Gelderman et al., 2006; Gunther and Bonnacorsi, 1996; Roodhooft and van der Abeelle, 2006).

Public procurement activities by European Union member states are based on the principles from the Treaty of Rome (1957) aiming for establishing the free market within the EU is the base for public procurement and follows five fundamental principles:

Non-discrimination – all discrimination based on nationality or by giving preferences to local companies is prohibited.

Equal treatment – all suppliers involved in a procurement procedure must be treated equally.

Transparency – the procurement process must be characterized by predictability and openness.

Proportionality – the qualification requirements must have a natural relation to the supplies, services or works that are being procured.

Mutual recognition – the documents and certificates issued by the appropriate authority in a member state must be accepted in the other member states.

A considerable part of all purchasing activities on any national market is due to public procurement and corresponds to around 17% of the European Union's GDP,.. *at the local/regional level public procurement can easily reach the double of that in terms of percentage of public expenditure. For the example, a study of public procurement across Baltic city metropolises(3) shows that public procurement accounts for 40% of the city budget in Helsinki and 30% in Stockholm.* (CORDIS – Community Research and development of Information Service, http://cordis.europa.eu/fp7/ict/pcp/key_en.html)

Knowledge Management combined with electronic way to carry out procurement activities can play a great role in increasing the efficiency of the whole process.

2. Method

The method we apply to study the procurement process in the light of knowledge flow is those of bottom-up supervised by a top down (Mercier-Laurent, 2007). It consists in studying first the process of procurement from activity, involved actors and a related flow of knowledge points of view. It includes internal and external sources of knowledge. The KADS[1] way of thinking help in this analysis – what is the goal, what are the knowledge involved in solving the given problem, what is the context (external knowledge, legal, environmental…) and what actions we suggest to solve a given problem (Breuker, 1997).

We also apply the needs engineering (Mercier-Laurent 2007 and 2011) to know and discover the unexpressed needs. It consists in observing the users activity and working with them on the future system functionalities. While the most actors focus on expressed needs, the unexpressed needs are the base of innovation.

The analysis of a knowledge flow will be a base of future organization of knowledge to support the procurement activity in connection with the others organization processes.

The KNUT project results are our starting point.

[1] Knowledge Acquisision Design Systems – European project Esprit 2

3. Electronic Public Procurement

Electronic procurement has over a number of years gained recognition in both private and public sectors. More and more parts of the procurement process are done electronically, even to carry out the whole tendering process electronically. Leukel and Maniatopoulos (2005) define *in a public sector context, e-Procurement* as *a collective term for a range of different technologies that can be used to automate the internal and external processes associated with the sourcing and ordering process of goods and services*.

In Sweden, in December 2005, Verva[2] published a report for a national action plan of procurement stressing that both buyers and suppliers would benefit from electronic methods for procurement and that the whole procurement process from planning to billing should and could be done electronically. Sveriges Kommuner och Landsting (The Swedish Association of Local Authorities and Regions) is heading another important Swedish project - SFTI (Single Face to Industry) that aims to standardise the communication between public and private organisations. It has received international attention.

The Swedish research project KNUT (Electronic Public Procurement of Telecommunications Services), sponsored by the Swedish government agency Vinnova[3], was an attempt to develop electronically the missing parts of the whole public procurement process by in a systematic and structured way collect and analyze needs and incorporate the legacy system. This approach is of special importance in the public procurement environment, since after the publishing of the RfP, no changes can take place. When the pre-formal phases of the public procurement process are carried out electronically, the results can be directly transferred to the electronically produced RfP. One of the objectives of the project was to increase the number of SMEs to participate in the public procurements. (Lindskog, 2010)

The first parts of the procurement process to be standardized and electronically applied were ordering and billing. This is due to the relatively low complexity and limited number of solutions for these parts of the procurement process.

According to IDABC - **I**nteroperable **D**elivery of European eGovernment Services to public **Administrations**, **Businesses** and **Citizens** (2009) eProcurement can benefit:

Public Administrations: eProcurement should minimize the time and effort expended by administrations and contracting authorities for organizing public procurement competitions.

***Businesses**: It will also benefit enterprises keen to trade across borders, by giving them improved and easier access to public procurement opportunities across Europe.*

***Citizens**: Making procurement procedures available to a larger audience of suppliers enables the public sector to purchase goods and services at more economically advantageous prices. Citizens will have reassurance that their administrations are spending money in a more cost-effective manner.*

[2] Verva was one of the Swedish Government's central advisory agencies. Today, public coordination of framework contracts concerning products and services for the entire public sector in the fields of information and communications is carried out by Kammarkollegiet.

[3] Vinnova - (Swedish Governmental Agency for Innovation Systems) is a State authority that aims to promote growth and prosperity throughout Sweden. The particular area of responsibility comprises innovations linked to research and development. The tasks are to fund needs-driven research required by a competitive business and industrial sector and a flourishing society, and to strengthen the networks that are necessary parts of this work.

The implications of electronic public procurement and the objectives for using electronic public procurement are many and can be summarized as:

- To reduce public sector spending

In the context of public procurement, cost reduction can be divided into three types:

  - Overall

    The reduction can be achieved by purchasing goods and/or services that correspond to the authorities' specific needs (not too much, which could unnecessarily increase costs, or too little, which risks not meeting the authority's current and future needs) or by changing internal processes using new technologies or other innovative approaches. One example can be increased usage of information communication technology, which may signify increased cost of purchasing in order to reduce the overall cost.

  - For specific service and/or goods

    Standardized procedures and descriptions of products reduce the cost for the preparation of tenders. Electronic procurement makes it possible for more suppliers to be aware of and easier bid for government contract and/or through economy of scale in case of framework contract

  - Of the procurement process

    By using models, standardized procedures for every phase of the public procurement process and concentrating on analysis of needs specific for the authority, thus, to reduce the amount of time and the number of staff involved in the process.

- To increase the service level

Usage of the same standardized procedures and description of goods and services makes it easier for the suppliers to know how to bid for the government contracts. Usage of a model for the analysis of needs and standardized Request for Proposal reduces the number of misunderstandings.

- To increase the number of SMEs
- To reduce the dependency on consultants

Electronic procurement gives the possibility for authorities to use the aggregate knowledge from earlier procurements in the same area and of similar type of authorities, thus, to reduce the need for external workforces, especially for less qualified type of information.

- To increase competition

Standardized procedures, models for structuring needs and requirements, and easier access to information about forthcoming public procurements and to electronic RfP facilitate the possibility for a bigger number of companies and from more EU countries to participate in public procurements

- To rationalize and increase efficiency

This is especially valid in case of carrying out the whole procurement process electronically. Even a partially electronic process such as for example eInvoicing can give substantial gains.

- To increase the possibility to aggregate knowledge

Electronic procurement makes it much easier to collect statistics over existing procurements and analyze criteria, requirements, contracts and/or outcomes in order to attune new specifications and to avoid mistakes.

- To increase the possibility to put pressure towards more standardization

If the statistics based on a big number of procurements show similar difficulties in achieving satisfactory results depending on the lack of standards for specific functionality, there is pressure on standardization organizations to develop and prioritize standards that are in demand by the users.

- To empower each agency

Easy access to information about how to make an analysis of needs using appropriate models, specify requirements automatically based on analysis of needs and develop a Request for Proposal corresponding to an authority's needs gives the possibility to carry out the entire procurement process by the agency itself.

The most important conclusion is that carrying out the whole procurement process electronically can lead to considerable gains of rationalisation and efficiency. The upgradeable and scalable model enables analysis of needs, collection of information about the legacy of currently valid contracts and a direct transformation of these results into the RfP. Thus, the whole process can be carried out electronically.

4. Analysis of a Public Agency purchasing activity

The purchasing activity of a public agency is presented in Figure 1.

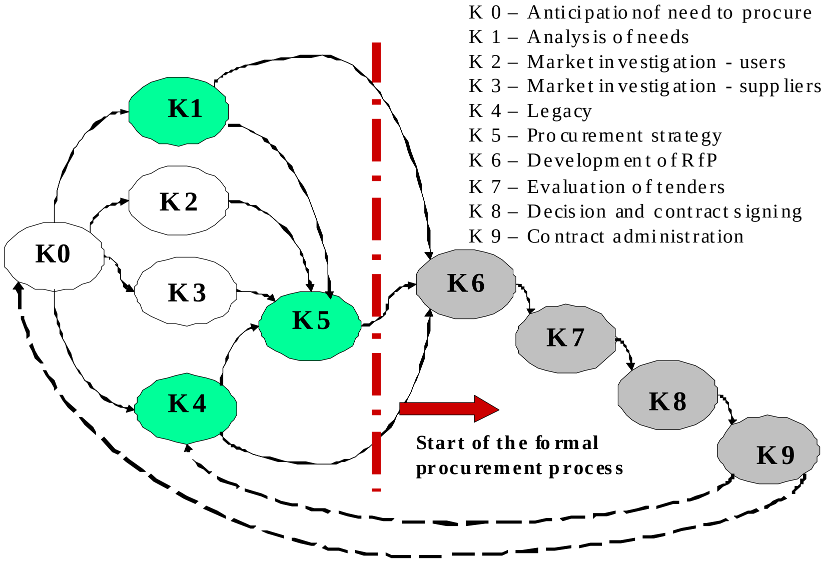

K0 – Anticipation of need to procure
K1 – Analysis of needs
K2 – Market investigation - users
K3 – Market investigation - suppliers
K4 – Legacy
K5 – Procurement strategy
K6 – Development of RfP
K7 – Evaluation of tenders
K8 – Decision and contract signing
K9 – Contract administration
K1
K2
K0
K3
K5
K4
K6
K7
K8
K9
Start of the formal procurement process

Figure 1 Public procurement – the purchasing process of a public agency (Lindskog, 2008)

The organizational buying process has been analyzed and structured by several researchers, among them Webster (1965), Robinson, Faris och Wind (1967), Wind and Thomas (1980), and Kotler (1997). Their research findings constitute the base for structuring and analysing the public procurement process of telecommunications services in the KNUT-project (Lindskog, 2010).

The public agency's buying process shown in Figure 1 is composed of ten phases and some of them can be carried out in parallel. The parts K1, K4 and K5 were developed in the KNUT project and K6, K7, K8 and K9 are the parts already available on the market.

K0 – Anticipation of need to start a procurement process. At this stage the organization, and especially its procurement department, observes the need of a new procurement. The absolutely most common reason to anticipate the need of the new procurement is the situation when the current contract is about to expire. If this happen close to the expiration date of the valid contract, it can be difficult to allocate enough time to carry out all necessary steps such as market investigations, analysis of needs, survey of legacy or choosing the procurement strategy.

K1 – Collection and analysis of needs.
This is an internal activity in order to know what is needed in detail. Collection and analysis of needs often start with an analysis of the current situation and sometimes with a formulation of vision and strategy to achieve the vision. The vision can concentrate on "core" activities, improvement of the service level towards citizens and businesses, increased efficiency and reduction of costs.

K2 – Market investigation – users.
This is an activity in order to find information about what others already have done in similar types of procurements. To meet other public agencies, private companies and/or users' associations in the own country or abroad and learn from their experiences from procurement of telecom services can be a very efficient way to develop RfPs and to avoid repeating errors committed by others. In contrast with private companies, there is no competition between public agencies, which gives possibilities to exchange experiences regarding suppliers, procurement process and internal difficulties. This input can be very valuable for the procuring agency in order to avoid problems or at least to be conscious of their existence.

K3 – Market investigation – suppliers
Public agencies are allowed to have contacts with manufacturers, operators and standardization organizations in the pre-study and market investigation phase. It is important to make use of this possibility in order to avoid unrealistic or costly requirements as well as to avoid missing important "in the pipeline" future services, solutions or functions.

K4 – Collection of information regarding legacy
This is an internal activity that investigates and collects information about already existing contracts and equipment within the organization. The most important parts of this investigation in case of telecommunications are legacies in form of ownership of properties, PABX[4]s, terminal equipment, and routers, and own networks such as building wiring or municipal broadband network. Other important parts are currently valid contracts on fixed connections, telephony services, switched board operators services, call center services etc. All these aspects must be taken into account in the development of the RfP.

K5 – Choice of procurement strategy

[4] PABX – Private Automatic Branch Exchange

In this activity, the procuring organization investigates possible procurement scenarios and carries out the analysis of the consequences for each of these scenarios. The choice of the procurement scenario heavily depends on earlier undertaken investigations in K1, K2, K3 and K4.
In case of procurement of telecommunications, the KNUT project found three main scenarios:

1. Purchase of equipment,
2. Leasing of equipment
3. Service procurement (procurement of function)

Each of the main scenarios has several sub scenarios.

K6 – Development of Request for Proposal (RfP)
The development of the RfP is the central internal activity for public procurement. It includes structuring of mandatory and non-mandatory requirements, decision upon evaluation criteria, and often also contract proposal. Phases K1, K4 and K5 are input values for this activity. In the case of well carried out analysis of needs, legacy and choice of procurement strategy, the development of the RfP can be done automatically, i.e. electronically. With the development of the RfP starts the formal procurement process. The RfP cannot be changed after being published.

K7 – Evaluation of tenders
Tenders that do not comply with mandatory requirements are rejected and most of the evaluation will be concentrated on non-mandatory requirements and prices following the evaluation criteria. As a result one or several suppliers are chosen for decision taking.

K8 - Decision taking and contract signing
Decision taken by the procuring organization is valid only after giving during the stipulated time the possibility for the loosing tenderers to make a court appeal if they consider themselves being mistreated.

K9 – Ordering and invoicing, and follow-up – After the contract is signed and up and running, the delivery and invoicing period starts depending on the type of goods or services. In case of framework contracts from a designated agency that procures on behalf of other public agencies, it is necessary to have a call-off contract with each specific agency that is calling off from the framework contract. The delivery and invoicing is to the calling-off agency. In order to learn from the specific procurement, both buyer and supplier should measure customer satisfaction and results/profits. This is an important and valuable input for decision making for tendering in other procurements in the same area.

The KNUT project aimed especially on the development of a model and a tool for the phases K1, K4 and K5 since the information from these phases can directly be transferred to already existing electronic procurement tools for the development of the RfP. The KNUT project developed the methodology and a tool for purchasing telecommunications services. The results of the project were tested in a real life procurement of telecommunications services in Swedish local community Lindesberg.

The experiences from this procurement could be used for procurements of telecommunications services by other entities and/or as a model for development of similar applications for other complex procurements. Possibly the most important learnt lesson from this the project and the test is that carrying out of the phases (K0 – K5) before the formal procurement phase starts are crucial in order to achieve the best results and it is also in these phases that the usage of Knowledge Management methods can of great help.

5. The role of Knowledge Management in Electronic Public Procurement

Knowledge Management (KM) has been introduced as a new management method in the late 1980s (Drucker 1992, Savage, 1990, Amidon, 1997) and a decade latter via ICT, mainly without taking into account artificial intelligence approaches and techniques. While one of the objectives of Artificial Intelligence (AI) has been always knowledge modelling and processing, the KM have been introduced in early 1990s via Corporate Memory concept (CEDIAG, 1991). Corporate Knowledge goal is to build an optimized knowledge flow for a given organization using conceptual knowledge modelling and storage, intelligent access to knowledge (semantic navigation in the knowledge repository) and make it available for decision taking (Mercier-Laurent). KADS[5], conceived for designing the knowledge-based systems (Dolenc, 1996) takes into account the contextual knowledge, which is, in many cases, essential in decision taking process. The importance of external knowledge of stakeholders, including clients (Amidon, 1997, Mercier-Laurent, 2011) for the organizational strategy has been understood in the same period. The definion of Knowledge Management we use is following:
The organized and optimized system of initiatives, methods and tools designed to create an optimal flow of knowledge within and throughout an extended enterprise to ensure stakeholders success (Amidon, 1997, Mercier-Laurent, 1997*).*
According to this definition, AI plays an essential role in the optimizing of the knowledge flow and processing the knowledge elements influencing the success of all participants.

In the case of public procurement process, the related internal knowledge contains the elements such as needs, potential suppliers, technical knowledge, available technology, past projects and experience.
The elements of external knowledge are legal, political, environmental, economic and technological nature.
The electronic procurement allow reducing paper, but also introducing the new way of thinking relative to usage of this media and instantaneous access to world base of information. Instead of reproducing the paper mechanism, it allows innovating in the way of preparing, diffusing and processing the proposals as well as in the way of capturing the opportunity and decision taking.

5.1. Knowledge flow for e-procurement

Considering the process presented in Figure 1, each step needs knowledge to act. Some elements of knowledge are shared by several steps. The steps organization could be considered as a flow of knowledge. The process of procurement exchange knowledge with other organizational processes, such as services provided to citizens and/or enterprises (for example tax collection, healthcare, car registration and so on) and practically all internal processes.

[5] Knowledge Acquisition Design Systems, European project

A preliminary analysis of a knowledge flow related to e-procurement process is presented in Figure 2

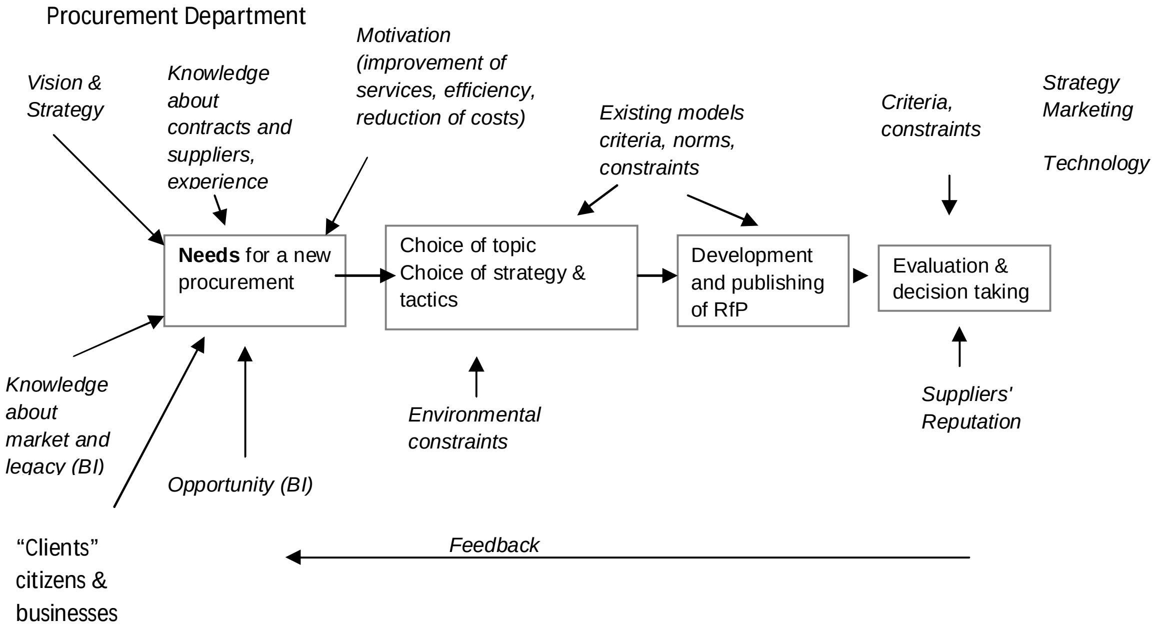

Procurement Department
Vision & Strategy
Knowledge about contracts and suppliers, experience
Motivation (improvement of services, efficiency, reduction of costs)
Existing models criteria, norms, constraints
Criteria, constraints
Strategy Marketing
Technology
Needs for a new procurement
Choice of topic Choice of strategy & tactics
Development and publishing of RfP
Evaluation & decision taking
Knowledge about market and legacy (BI)
Opportunity (BI)
Environmental constraints
Suppliers' Reputation
"Clients" citizens & businesses
Feedback

Figure 2 Knowledge flow for procurement process

The anticipation and definition of needs involve the elements as knowledge on existing contracts and constraints, vision, strategy and related tactics. At this stage the business intelligence (BI) techniques (text mining, semantic search…) can be helpful. BI is also useful in continuous discovering of new opportunities. Collection and analysis of needs can be improved by adding "needs discovery" aspects (Mercier-Laurent, 2007). It includes the involvement of selected users into the process. In function of a type of organization it may include the knowledge about citizens and businesses.

Market investigation involves knowledge on similar types of procurement, about experiences, users and suppliers. This knowledge can enhance the effectiveness of the whole process. A system of practice collection and processing, using for ex analogy could be an added value. The opening to external knowledge, such as experience with suppliers and tools is vital.

The information on legacy and other constraints to be considered can be collected by all participants and semantic search.

Choice of procurement strategy and tactics will be based on knowledge of benefits related to the various scenarios (purchase, leasing or function procurement). Corporate Social Responsibility (CSR) and norms encourage choosing rather the last scenario, if possible. The benefits and impact of a given choice can be simulated to help decision taking.

Knowledge modeling techniques can help in quicker RfP construction. The phase of proposals evaluation may use a decision support system and constraint programming.

The described flow is the first step of Knowledge Management approach. It should be improved by working on KNUT project case

6. Conclusions and perspectives

Public procurement can be a first step in introducing a knowledge management approach in organizations (bottom-up method). The described case is just a beginning of a larger project which could be followed by a PhD student working on a real case.

The Knowledge Management approach will increase the efficiency of the procurement process by organizing and optimizing related knowledge collection, modeling and processing. The reuse of existing knowledge components and experiences, as well as the quick access to relevant information will help the quicker development of RfP, a decision support system will assist in the process of evaluation and choice of supplier. AI approaches and techniques can bring a significant help in knowledge gathering, modeling, reusing as well as in discovering knowledge from data or text. Concerning SMEs the whole procedure should be simplified to focus on essential. A SME can not spend a month answering if it is not sure to gain the contract.

The integration of Corporate Social Responsibility will bring a feedback to e-procurement policies and process.

7. References

Ahlström, M. (2000). *Offset Management for Large Systems – A Multibusiness Marketing Activity*. Linköping University, Studies in Management and Economics,

Amidon D. (1997) *The Innovation Strategy fot the Knowledge Economics*, Butterworth Heinemann, ISBN 0750698411

Bonaccorsi.A., Pammolli, F. & Tani, S. (1996). *The changing boundaries of system companies*. International Business Review, 5 (1);

Breuker G.Schreiber, B.Wielinga, J.Breuker (1993): *KADS, A Principled Approach to Knowledge-Based System Development*, Academic Press 1993

CEDIAG Corporate Knowledge (1991) internal Bull document (confidential), described in Mercier-Laurent, 2004

Coase, R. (1937). *The Nature of the Firm*, Economica 4 (16): 386–405,

Commission of the European Communitites (2006). *Public Procurement: Legislation*, Direction General XV, Bruxelles. available at: http://ec.europa.eu/youreurope/nav/en/business/public-procurement/info/legislation/index_en.html

Drucker P. (1992) The New Society of Organizations, Harvard Business Review, September-October 1992

European Commission (2004) *Directive 2004/18/EC of the European Parliament and of the Council of 31 March 2004 on the coordination of procedures for the award of public works contracts, public supply contracts and public service contracts*

Gelderman, C.J., Ghijsen, P.W.T. & Brugman, M.J. (2006). *Public procurement and EU tendering directives – explaining non-compliance.* International Journal of Public Sector Management, 19 (7), 702-714.

Goldkuhl, G and Axelsson, K (2007) – *E-services in public administration*, International Journal of Public Information Systems, vol.2007:3, pp. 113-116

Kearney, A.T (2005) *Public Procurement – Achieving Value for Money whilst Enhancing Competition*, ATKearney Ltd

Karlsson, Magnus (1998) - *The Liberalisation of Telecommunications in Sweden Technology and Regime Change from the 1960s to 1993*, Linköping University, Tema T, Linköping,

Kotler, Philip (1997) - *Marketing Management Analysis, Planning, Implementation, and Control*, 9th edition, Prentice Hall, USA,

Leukel, J., & Maniatopoulos, G. (2005) - *A comparative analysis of product classification in public vs. private e-procurement. The Electronic Journal of e-Government, 3*(4), 201-212

Lindskog, H. (2004). *How can the private sector benefit from the public sector's e-procurement experiences?* in Morgan, K. et al (eds.) The Internet Society: Advances in learning, Commerce and Security. WIT Press, 123-138

Lindskog, H. (2005). *SOTIP as a Model for Outsourcing of Telecom Services for the Public Sector.* proceedings from HICSS conference, Hawaii, US, Lindskog, H. & Johansson, M. (2005). *Broadband: a municipal information platform: Swedish experience.* International Journal of Technology Management, 31 (1/2) 47-63

Lindskog, H. **(**2005) - *SOTIP as a Model for Outsourcing of Telecom Services for the Public Sector*, Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 8 - Volume 08,

Lindskog, H., Bjurman, Pernilla (2005) - *Public Procurement of Telecom Services*, ITIB conference, St Petersburg,

Lindskog H. (2006) *Ethical considerations of Outsourcing of Call Center Function in the Public Sector.* Proceedings of 9th QMOD (Quality Management and Organisational Development) Liverpool

Lindskog, H. (2008) – *Process of Public Procurement of Telecom Services, The Buyer's Perspective,* Proceedings of the 7th ISOneWorld conference, Las Vegas

Lindskog H., Brege & Brehmer (2008). *Private and public sectors' purchasing. How do they differ?.* proceedings from MODEST conference, Rembertow

Lindskog, H (2010). *KNUT Elektronisk offentlig upphandling av telekommunkationstjänster.* slurapport, Vinnova

McAdam, R. & McCreedy, S. (2000). *A Critique of Knowledge Management: Using a Social Constructionist Model.* New Technology, Work and Employment 15

Mercier-Laurent E. (1997) *Global Knowledge Management beginning from website - How to organize the Flow of Knowledge in an International Company -theories and practice, ISMICK 97, Compiegne*

Mercier-Laurent E. (2004) From *Data Programming to Intelligent Knowledge Processors: How Computers Can Improve the Global KM Flow*, Cutter IT Journal, Vol. 17, N° 12, December 2004 p. 33-39

Mercier-Laurent E. (2007), *Rôle de l'ordinateur dans le processus global de l'innovation à partir de connaissances*, Mémoire d'Habilitation à Diriger les Recherches en Informatique, Université Lyon3.

Mercier-Laurent E. (2011) Innovation Ecosystems Willey-ISTE, ISBN: 978-1-84821-325-8,

Nonaka, I (1991). *The knowledge creating company.* Harvard Business Review 69; 96-104

Olson, E.L, & Bakke, G. (2001). *Implementing the lead user method in a high technology firm: A longitudinal study of intentions versus actions.* The Journal of Product Innovation Management, 18, 388-395.

Prahalad, C.K. & Hamel, G. (1990). T*he Core Competence of the Corporation*, Harvard Business Review, May-June

Robinson, P.J., Faris C.W. & Wind Y. (1967). I*ndustrial Buying and Creative Marketing*. Allyn and Bacon Inc., Boston.

Roodhooft, F., & Van den Abbeele, A., (2006). *Public procurement of consulting services: Evidence and comparison with private companies.* International Journal of Public Sector Management, 19 (5), 490-512.

Savage C. (1990) 5th Generation Management: Integrating Enterprises through Human Networking, The Digital Press, Bedford

Strauss, A and Corbin, J (1998) – *Basics of qualitative research techniques and procedures for developing grounded theory*, 2nd edition, Sage, Newbury Park,

Dolenc N., Libralesso J-M, Thirion C., Gobrecht A., Lesaffre F-M, Hellelsen M., Lalhier M., Lenuet D., Steller J-M. (1996) *The SACHEM Project : Blast Furnace Operating Support System ; Ambition and Stakes, Development and First Results* 3rd European Ironmaking Congress, Gent, Belgium Sept 16-18 1996

Thompson, M. P. A. & Walsham, G. (2004). *Placing Knowledge Management in Context.* Journal of Management Studies 41 (5): 725–74

Treaty of Rome – 25 March 1957

von Hippel, E. (1985). *Learning from Lead Users.* in Buzzell R.D (ed) Marketing in an Electronic Age, Boston: Harvard Business School Press, 308-317

Williamsson, O.E. (1975). *Markets and Hierarchies: Analysis and Antitrust Implications.* The Free Press, New York

Williamsson, O.E. (1979). *Transaction-cost economies: the governance of contractual relations.* The Journal of Law and Economics, Vol.22 No.3, 233-261

Williamsson, O.E. (1981). *The Economics of Organization: The Transaction Cost Approach*. The American Journal of Sociology, 87(3), 548-577

Williamsson, O.E. (1985). T*he Economic Institutions of Capitalism*,. The Free Press, New York

Wynstra, J.Y.F. (1998). *Purchasing involvement in product development*. Doctoral thesis, Eindhoven Centre for Innovation Studies, Eindhoven University of Technology, 1998

Webster F.E. (1965) – *Modeling the Industrial Buying Process*, Journal of Marketing Research, Vol. 2, pp 370-6

Wind, Y. and Thomas, J. R. (1980) – *Conceptual and Methodological Issue in Organizational Buying Behaviour,* European Journal of Marketing, vol.14, pp 239-286

http://www.eurodyn.com/

http://ec.europa.eu/idabc

http://europa.eu/publicprocurement

http://osor.eu

http://www.peppol.eu

http://ted.europa.eu

http://cordis.europa.eu/fp7/ict/pcp/key_en.html

Collective intelligence for Evaluating Synergy in Collaborative Innovation

Ayca Altay and **Gulgun Kayakutlu** [1]

Abstract. Collaborative innovation is an unavoidable need for the small and medium enterprises (SME) both in terms of economic scale and technological knowledge. Risks and the innovation power are analyzed for the wealth of collaboration. This paper aims to present the *synergy index* as a multiplier of the innovation power of research partners to make the collaboration successful. The proposed index can be used with different number of companies in collaboration cluster and the synergy maximization is guaranteed by using a new particle swarm algorithm, *Foraging Search.* This paper will give the formulation and criteria of the synergy index in detail. A sample synergy index application for the Turkish SMEs will clarify the steps to follow.

1 INTRODUCTION

Recently an article published in Scientific American illuminated one of the main differences between the humans and the animals as sharing the knowledge to create cumulative culture [1]. Though it is recently biologically proven, we have been using the concept of synergy in engineering since the very first project developed to create a team work. In the last few years international projects are run by in collaboration by public and private authorities causing studies and discussions on synergy and conflict [2]. Companies are obliged to innovate for competition and are willing to collaborate for the unique product/processes/service only after defining the team with bigger chance of success [3]. Small and medium companies (SME) would like to gratify the collaborative innovation with less risk.

The main approach in the synergy literature is the extraction of factors that affect synergy in alliances using case studies [4][5] or statistical analyses [6][7]. These studies recommend building alliances based on the criteria that have the biggest effect on collaborations. Further quantifications are achieved with Multi Criteria Decision Making, where partners are selected using the criteria extracted in previous studies [8]. In the existence of strict goals of alliances, a number of mathematical methods are built. Majority of the researchers have exploited and developed recent mathematical models involving the goal programming [9] and multi-objective programming [10].

This study has the main objective of proposing a synergy index as a multiplier of the innovative power to be maximized for successful partnership. It will be presented that when both innovation capabilities and the risks are considered through internal and external influencers, the synergy created will avoid the failure. In order to determine the best team of companies, the possible companies are to be clustered based on all the criteria effective for synergy improvement. A collective intelligence approach, particle swarm optimization is selected to evaluate the collaborative synergy since it has the social component in parallel with the knowledge based evaluation [11]. However, the fact that the classical particle swarm method is based on balancing the exploration and exploitation at the particle level [12] would mean individual success of each company. An advanced new particle swarm algorithm foraging search is based on creating balance of exploitation and exploration at the swarm level as well as particle level, which allows us to calculate the collaborative success [13].

This paper is distinguished and will make contribution to the research in three main points:

- Instead of choosing a partner as studied before, this study deals with grouping and clustering of the synergy creating SMEs.
- The criteria studied in this research combines the innovative power and risk criteria with the synergy which are depicted from the literature and selected by industrial experts.
- Algorithm used in this study is not based on a threshold as in goal programming, thus it allows the selection of partners even in vague and uncertain conditions.

This paper is so organized that a literature review on the collaborative innovation will be given in the next section and the synergy index function will be explained in the third section. Foraging Search algorithm that is used to maximize the synergy will be explained in the fourth section and the fifth section of the paper is reserved for the application. The conclusion and further suggestions will be summarized in the last section.

2 SYNERGY IN COLLABORATIVE INNOVATION

Knowledge based collaboration is the fuel of innovation for the SMEs. They are known to be agile in change, but fragile in facing the economic fluctuations [14]. Collaborative innovation is mainly based on the synergy created by the partner companies. When it is on the virtual network an intelligent agent can take the role of a moderator. In private or public industries skill based clustering has been an effective tool to create synergy among the team workers [15][16]. But, it is difficult to construct a creative task ground for the team members who come from different business cultures. Innovative capabilities of more than one company working together are established on both the knowledge and vision for

[1] Industrial Engineering Department, Istanbul Technical University, Macka 34367 Istanbul, Turkey email: kayakutlu@itu.edu.tr

internal and external alliance. Big companies succeed the collaboration by defining the performance focused on cross-business growth [17]. They might even improve the innovative capabilities by merger and acquisition [18]. SMEs on the other hand, see the research support as one of the external fund to be accessed [19] and they jump into any partnership even it might be quite risky. Chang and Hsu studied both managerial and environmental drivers of innovativeness for SMEs to show that internal and external factors are independent [20]. Global collaboration changed the collaborative strategies both in functional operations and collaborative activities [21]. The economic crisis has led research and development for innovation towards a new approach and perspective: innovation through new products, processes and knowledge is not enough beneficial unless the systems around them are not ready. This is a common issue among the developing and highly developed countries [22].

Literature surveys allowed us depict forty-four innovative synergy factors representing either organizational approach (summarized in Table 1) or alliance approach (summarized in Table 2).

Previous research also shows that these criteria are mainly analyzed by constructing the clusters in the same geographical region by using the collective intelligence methods [23][24].

Table 1. Organizational Features Effective in Collaboration Synergy

Factor	Information Resource	Reference
Level of education	Diplomas/certificates	Chang&Hsu(2005)[20]
Organizational structure	Organizational Manual & Management Survey	Twardy (2009)[25]
Accounting procedures		Margoluis (2008)[26]
Compensation Policies & Procedures		Margoluis (2008)[26]
Performance culture	Management Survey	Rai et al.(1996)[27]
Governmental support	Country Regulations	Rai et al.(1996)[27] Ding (2009)[28]
Legal culture	Legal Assessment	Twardy (2009)[25]
Financial condition	Company Balance Sheet	Rameshan&Loo (1998)[29] Chen et al(2008)[30] Twardy (2009)[25] Ding (2009)[28]
Organizational resources	Company balance sheet & Management Survey	Margoulis(2008)[26]
Resources for R&D		Chen et. al(2008)[30]
Technological Capabilities	Technology Assessment	Chen et. al(2008)[30]
Geographical scope	Sales Information	Ding (2009)[28]
Unique Competencies	Number of Patents	Ding (2009)[28]
Visions, Goal&Objectives	Employee & collaborator Survey	Margoluis (2008)[26] Gomes-Casseres (2003) [31]
Type of Leadership		Margoulis(2008)[26]
Past cooperation experience		Chen et. al(2008)[30]
Future Expectations	Employee & Management Survey	Linder et. al.(2004)[32] Twardy (2009)[25]

Organizational factors also includes intangible factors like brand / firm reputation [26], inter-organizational trust, commitment capabilities to alliance, inter-organizational information share, company pace, common errors and leader attitude. Alliance vision embraces some intangible factors like commitment, experience sharing, sales & marketing coordination. These intangible factors can be evaluated through employee, management and collaborator surveys only.

The alliance approach also includes intangible criteria in addition to the tangible ones given in Table 2. The intangible factors consist of mutual trust, information and experience sharing.

Table 2. Alliance Features Effective in Collaboration Synergy

Factor	Information Resource	Reference
Scope, goals and objectives clarity	Contract, Employee Survey	Eden (2007)[34] Margoluis(2008)[26]
Structure of the alliance (clarity of roles)		Margoluis(2008)[26]
Division of labor		Margoluis(2008)[26]
Funding participation		Linder et. al.(2004)[32]
Project manager	Project Manager Profile	Rai et. Al.(1996) [27] Margoluis (2008)[26] Eden(2007)[34]
Cooperation strategies	Management Survey	Chen et al(2008)[30] Twardy (2009)[25]
Dysfunctional conflicts		Rameshan&Loo (1998)[29]

Some of those factors are very similar and most of them cannot be expressed in figures. They are combined into 23 criteria according to the similarities after a fuzzy cognitive survey responded by the SME managers [13].

3 Synergy Index

3.1 Synergy Index Formulation

Synergy is defined as the concept of generating a greater sum than the sum of individuals [35]. The better is the accordance within the alliance, the greater is the synergy. Hence, synergy is positively related with the accordance. In other words, the system that makes the alliance work has to be robust for a lifetime of an alliance. Reliability can be defined in good working synergy criteria when the expected life of collaboration is the concern. The expected lifetime of alliance can be calculated using Weibull distribution which is accepted as the best function of lifetime calculation in the reliability theory [36, 37]. Weibull distribution will be constructed for each company considering the synergy coefficients and the number of companies in collaboration as parameters. Inter-company synergy will assume to have more than one firm in alliance. Weibull distribution has the following features:

Density function : $f(x) = \frac{\beta}{\alpha}\left(\frac{x-v}{\alpha}\right)^{\beta-1} \exp\left\{-\left(\frac{x-v}{\alpha}\right)^{\beta}\right\}$ (1)

Cumulative function: $F(x) = 1 - \exp\left\{-\left(\frac{x-v}{\alpha}\right)^{\beta}\right\}$ (2)

where $x > v$ and α, β and v are Weibull parameters.

Expected value: $E[X] = \alpha.\Gamma(1+\frac{1}{\beta})$ (3)

The analogy between the synergy and the lifetime suggests $v \geq 0$, since we take the two, analogous $v = 0$ will be accepted. In the formula β is the shape parameter and α is the rate parameter. When β=0, the Weibull distribution becomes the Exponential distribution. In physical and biological systems, synergy is modeled with an accelerating effect, which resembles the shape of exponential distribution [37]. This allows us take the shape parameter β to denote the *number of firms in collaboration.*

The parameter α resembles the strength of elements in the reliability analogy, which is equivalent to the *merged synergy coefficient* that will be calculated using synergy factors.

The synergy index S can be defined as

$$S = \alpha.\Gamma(1+\frac{1}{\beta}) \qquad (4)$$

α: the merged synergy coefficient
β: number of companies

The synergy index will be used in calculating the maximization of innovation power. It is known that in collaborations the innovation can be greater than the sum of the individual if the accordance is well established. Hence, we try to maximize the minimum synergy among the collaborating companies. Each collaborating group is important and the company left outside the group must be successful if included in collaboration.

It should also be clear for the collaborating companies that if the synergy factors are merged in a negative way, that is, if the companies are discordant, the synergy index will be negative showing no possible lifetime for collaboration clusters.

3.2 Sensitivity Analysis

The proposed synergy index is sensitive to the number of firms in alliance. As an example, there exists 2 collaboration clusters, one with 2 companies and the second with 3 companies. In case the merged synergy coefficient α=0.7 for both clusters 3-company-alliance gives a better S than the 2-company alliance. This can be considered as a parallel system. It is always safer to increase the number of parallel elements. In Figure 1, synergy index sensitivity of number of firms in alliance for α = 0.7 is demonstrated.

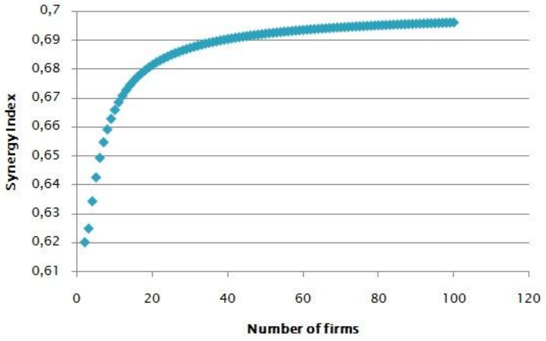


Figure 1. Synergy index sensitivity

Synergy effect in innovation is shown to support the moral as well as causing improvements in project follow-up, creativity and technological intelligence when thresholds are taken into account [38]. The previous evaluations of synergy were mainly based on scoring because of the intangible factors.

4 Foraging Search

4.1 Motivation

The Foraging Search algorithm imitates the Animal Food Chain for optimization problems [39]. Animal Food Chain contains three groups: herbivores (plant eaters), omnivores (both plant and meat eaters) and carnivores (meat eaters). Herbivores are known as primary consumers, omnivores who feed on some specific plants and other herbivores are known as secondary consumers and lastly, carnivores who feed on specific herbivores and carnivores are known as the tertiary consumers. Herbivores are ultimate hunts of the food chain whereas carnivores are the ultimate hunters and omnivores, which are both hunters and hunts. According to the energy transformation, the energy transmitted through a food chain decreases as the number of consumers increase. The ratio of hunt and hunters depend on the ecological environment. In wild environments, the herbivore-omnivore-carnivore ratio can be 10:3:1 whereas in calm environments the related ratio can be 40:10:1. Additionally, it is also valid that in a food chain, the hunter is always faster than the hunt [40][41].

The classical PSO algorithm employs one swarm and the related swarm is responsible for both exploration and exploitation [42]. There is a new algorithm that implements two swarms of equal sizes clustering [43] separating the responsibilities for exploration and exploitation.

The Foraging Search uses three swarms, namely herbivores, omnivores and carnivores, to provide exploration by the herbivore swarm, exploitation by the carnivore swarm and exploration-exploitation balance by the omnivore swarm. Introduction of a food chain provides an incremental escaping ability which is modeled with first level and second level hunters. All fear and escape factors affect the speed of the animals, which the Foraging Search model embeds in the velocity update formula. Furthermore, this algorithm considers the environmental wildness which represents the complexity of the market. If the competition is harsh it is better to increase the wildness. That is why Foraging Search balances the exploration and exploitation at the swarm level.

4.2 The Clustering Algorithm

Each particle in the Foraging Search Clustering algorithm is represented by *k*d* cluster centers where *k* is the number of clusters and *d* is the number of dimensions of the data points to be clustered. Likely, the velocity and speed updates are applied in order to locate optimum cluster centers.

The following steps are followed:

Step 1. The environment is defined as calm, regular or wild.

Step 2. The herbivore : omnivore : carnivore (*h_number:o_number:c_number*) ratio is determined.

- IF the environment is harsh: wild 10:3:1
- IF the environment is average: 25:6:1
- IF the environment is calm: 40:10:1

Step 3. Each particle is randomly initiated for each swarm, each particle is assigned random *c*d* cluster centers where *c* is the number of clusters and *d* is the dimension of data points. The particles are named as x_{ijk}, the k^{th} dimension of the j^{th} cluster of the i^{th} particle where i = 1, ... *h_number* $\vee$ *o_number* $\vee$ *c_number*, j = 1, ..., c, k = 1, ..., d.

Step 4. Data points are assigned to clusters using a distance metric (e.g. Euclidean distance, Mahalanobis distance, etc...).

Step 5. The quality of the clustering is measured by an objective function. The aim of clustering is building small clusters as dissimilar as possible. Consequently, the objective function may involve within cluster distances, among cluster distances or a combination of both measures.

Step 6. The best objective value and position for all particles, or particle bests, are determined for each particle in each swarm.

Step 7. The best objective value and position, or swarm bests are determined for each swarm.

Step 8. The best objective value and position of all swarms, or the global best is determined.

Step 9. The fear coefficients for herbivores are calculated as follows:
Fear factors for herbivores :

$$pfho_i = 1 - \frac{d_{fho,i}}{d_{fho}^{\min}} \tag{5}$$

$$pfhc_i = 1 - \frac{d_{fhc,i}}{d_{fhc}^{\min}} \tag{6}$$

where
$i = 1,\ldots,$ *h_number*
$pfho_i$: fear degree from omnivores of the i^{th} herbivore (in the interval [0,1])
$pfhc_i$: fear degree from carnivores of the i^{th} herbivore (in the interval [0,1])
$d_{fho,i}$: the distance of the i^{th} herbivore to the nearest omnivore
$d_{fhc,i}$: the distance of the i^{th} herbivore to the nearest carnivore
$d^{min}{}_{fho}$: the minimum distance for a herbivore to fear an omnivore
$d^{min}{}_{fhc}$: the minimum distance for a herbivore to fear an omnivore

Step 10. The fear coefficients for omnivores are calculated using the formula below:

$$pfoc_i = 1 - \frac{d_{foc,i}}{d_{foc}^{\min}} \tag{7}$$

where
$i = 1,\ldots,$*o_number*
$pfoc_i$: fear degree from carnivores of the i^{th} omnivore (in the interval [0,1])
$d_{fho,i}$: the distance of the i^{th} omnivore to the nearest carnivore
$d^{min}{}_{fhc}$: the minimum distance for an omnivore fear an omnivore

Step 11. The probability of being a hunt for omnivores is calculated as

$$pp_i = \frac{dc_i}{dc_i + dh_i} \tag{8}$$

where
$i = 1, \ldots,$ *o_number*
pp_i: : the probability of omnivores being a hunter
dh_i: the distance of i^{th} omnivore to the nearest herbivore
dc_i: the distance of i^{th} omnivore to the nearest carnivore

Step 12. The velocities (v_{ijk}) of each particle are updated according to their swarms.

a. *Velocity Update for the Herbivore Swarm*
Since herbivores are ultimate hunt, their velocity update involves the escape from their first and second level hunters: omnivores and carnivores. The velocity update formula for herbivores is given below.

$$v_{ijk} \leftarrow wv_{ijk} + c_1 r_{1i}\left(y_{ijk} - x_{ijk}\right) + c_2 r_{2i}\left(\hat{y}_{ijk} - \right. \tag{8}$$

where
$i = 1,\ldots.,$ *h_number*
v_{ijk}: the velocity of k^{th} dimension of the j^{th} cluster of the i^{th} particle of the swarm
w: the inertia coefficient
$c_1$and c_2: cognitive and social coefficients
r_{1i}, r_{2i} ,r_{3i} and r_{4i}: random numbers for the i^{th} particle in the interval [0,1]
y_{ijk}: personal best for the k^{th} dimension of the j^{th} cluster of the i^{th} particle of the swarm
x_{ijk}: the position of k^{th} dimension of the j^{th} cluster of the i^{th} particle of the swarm
$\hat{y}_{ijk}$: swarm best for the k^{th} dimension of the j^{th} cluster of the i^{th} particle of the swarm
c_3: distance based coefficient of herbivores from omnivores
c_4: distance based coefficient of herbivores from carnivores
and $D(.)$ is a measure of the effect that the hunter has on the hunt and it is formulated as

$$D(x) = \alpha e^{-\beta x} \tag{9}$$

where d is the Euclidean distance between the prey particle and the nearest hunter particle. α and β are positive constants that define the effect of distance to velocity.

b. *Velocity Update for the Carnivore Swarm*
Since herbivores are ultimate hunt, their velocity update involves independently chasing the nearest hunt. The velocity update formula for herbivores is given below.

$$v_{ijk} \leftarrow \left(\hat{y}_{ijk} - x_{ijk}\right) \tag{10}$$

where
$i = 1,\ldots,$*c_number*
v_{ijk}: the velocity of k^{th} dimension of the j^{th} cluster center of the i^{th} particle of swarm
r: random number in the interval [0,1]
$\hat{y}_{ijk}$: the position of the k^{th} dimension of the j^{th} cluster center of the nearest hunt to the i^{th} particle of swarm
x_{ijk}: the position of the dimension of the j^{th} cluster center of the i^{th} particle of swarm

c. *Velocity Update for the Omnivore Swarm*
Since omnivores are both hunters and hunts, their velocity update involves the compound of both velocity update formulas whose ratio depend on the probability of being a hunter. The velocity update formula for herbivores is given below.

$$c_1 r_{1i}\left(y_{ijk} - x_{ijk}\right) + c_2 r_{2i}\left(\hat{y}_{ijk} - x_{ijk}\right) + pfho_i c_3 r_{3i} D\left(d_{fho,\right.}$$

$$v_{ijk} \leftarrow \left(1 - pp_i\right)\left(wv_{ijk} + c_1 r_{1i}\left(y_{ijk} - x_{ijk}\right) + c_2 r_{21}\left(\hat{y}_{ij}\right.\right. \tag{11}$$

where
$i = 1,\ldots.,$ *h_number*
v_{ijk}: the velocity of k^{th} dimension of the j^{th} cluster of the i^{th} particle of the swarm
w: the inertia coefficient
$c_1$and c_2: cognitive and social coefficients
r_{1i}, $r_{2i,}$ r: random numbers for the i^{th} particle in the interval [0,1]
y_{ijk}: personal best for the k^{th} dimension of the j^{th} cluster of the i^{th} particle of the swarm
x_{ijk}: the position of k^{th} dimension of the j^{th} cluster of the i^{th} particle of the swarm
$\hat{y}_{ijk}$: swarm best for the k^{th} dimension of the j^{th} cluster of the i^{th} particle of the swarm
$\tilde{y}_{ijk}$: the position of the k^{th} dimension of the j^{th} cluster center of the nearest hunt to the i^{th} particle of swarm

c_3: distance based coefficient of herbivores from omnivores
c_4: distance based coefficient of herbivores from carnivores
and $D(.)$ is a measure of the effect that the hunter has on the hunt and it is formulated as

$$D(x) = \alpha e^{-\beta x} \tag{12}$$

where d is the Euclidean distance between the prey particle and the nearest hunter particle. α and β are positive constants that define the effect of distance to velocity.

Step 13. The particle positions for each particle in each swarm are updated using the formula below:

$$x_{ijk} \leftarrow x_{ijk} + v_{ijk} \tag{13}$$

where
x_{ijk} : the position of k^{th} dimension of the j^{th} cluster of the i^{th} particle of the swarm
v_{ijk}: the velocity of k^{th} dimension of the j^{th} cluster of the i^{th} particle of the swarm

5 Case Study

The synergy index is studied for a case of 51 SME companies in Thrace, Turkey. The companies are distributed in several industries as shown in Table 3.

Table 3. Industrial distribution companies in the case study

Industry	%	Industry	%
Food	17.6	Service	15.7
Clothing & Textile	13.7	Health	5.7
Machinery & Electronics	19.6	IT & Communication	7.8
Automotive	2.0	Construction	2.0
Chemical & Pharmaceutical	0.0	Furniture	2.0
Plastics	2.0	Metal	2.0
Publishing	2.0	Miscellaneous	7.9

A survey of 23 questions is run for 51 companies to figure out the approaches on 23 synergy factors. Responses are clustered using the Foraging Search Algorithm and the synergy is calculated in clusters. The case is run by modifying the number of clusters from 2-25 and the results are shown in Figure 2. The objective function is maximizing the minimum synergies in the clusters. Hence, each cluster is important and should not be dispersed.

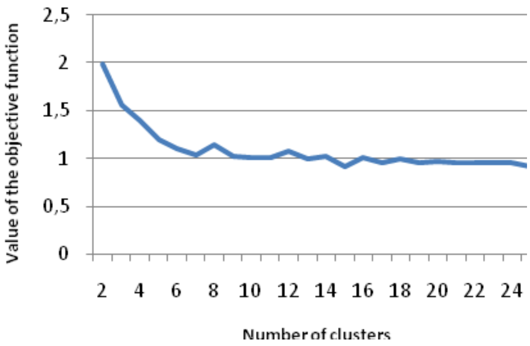


Figure 2. The synergy index with respect to the cluster numbers 2-24

Even if a company is left outside the clusters, it cannot be taken into any group without ensuring that it will be successful in collaboration.
Synergy index [illegible] for one company being equal to 1 shows that the company keeps its self innovation power as is in the collaboration. The cumulative estimated function is exp ([illegible]) is expected to be in the range (-1,1), so exp ([illegible]) being 0 means there is no synergy effect in the collaboration at all while -1 is negative and +1 is the positive effect of synergy.
As observed in Figure1, the 2 cluster trial gives the best value for the objective function with 1.98. There are 29 companies in the first cluster and 22 in the second cluster. The best value is obtained by the SME 1 and SME 2, which have a minimum 98% innovative synergy among them. The fact that several runs give the same exact solution makes us believe this is the global optimum.

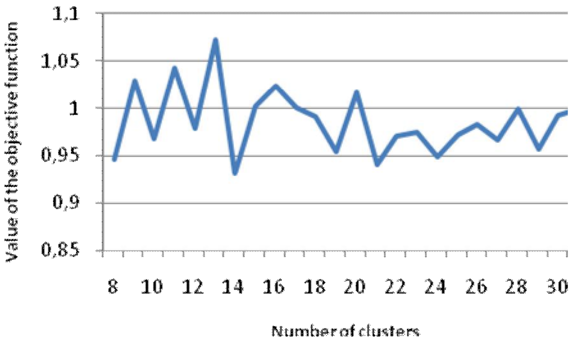


Figure 3. The synergy index with respect to the cluster numbers 8-33

Since the clusters are overcrowded in two clusters, another case is run to change the number of clusters between 8 and 33 as shown in Figure 3. The best objective value obtained is 1.04 with 11 clusters each having number of companies {5},{5},{3},{3},{4},{3},{4},{2},{ 5},{7} and {10}. Minimum synergy effect is obtained to improve 4%. This means the life of collaboration is prolongated from 1 year to 1.04 years. The eleventh cluster only includes SME 2 from the 2 cluster trial as a successful company. Cluster 11 owns ten companies from different industries in variety of sizes as shown in Table 4.

Table 4. Main features of companies in most successful cluster

SME No	Industry	Size
2	Electronics	Micro
3	Courier	Micro
4	Security Service	Micro
5	Electronics	Micro
7	Cable	Small
8	Textile	Micro
9	Security Service	Micro
27	Steel Production	Medium
37	Electronics	Medium
40	Auto-Spare Part Service	Small

It is observed that the successful collaboration is foreseen among companies from the most classical industries like textile and steel production and most technological industries like electronics and security service. Table 4 also shows that micro, small and medium companies can work together. Hence we can conclude that, unlike a generalized belief of industry and technology focus in collaboration, synergy is not only based on industry. All twenty-three factors are evaluated by the respondent companies and the most critical influencers are evaluated as alliance approaches and balance of the resources. Since a micro-firm cannot invest in research and development as much as a medium size company the focus is more human resource oriented.

6 Conclusion and Suggestions

Innovative synergy is requested for collaborative research and development that is an obligatory process for the small and medium companies. This study proposes a synergy index that will help the SMEs to decide which companies will maximize the synergy if collaborated. The synergy is accepted as the life of collaboration which will be prolonged with robust partnerships.

A case study among the 51 SMEs in Thrace, Turkey showed that the synergy is maximized with increasing number of companies. 98% synergy is obtained with 29 companies in a group. The real life shows that it might be interesting for cooperative activities like commerce chambers but not feasible for research and development or innovation. When number of clusters is increased to 11, the best synergy is obtained with a group of ten companies. It is observed that the highest innovation is received with companies with different sizes and from a variety of industries. The business and alliance approaches of companies have a bigger role in synergy. This conclusion suggests that SMEs are to be trained to collaborate with the companies that strengthen their weak points.

The proposed approach is to be further developed by validity analysis through comparisons with different approaches and different methods. It is also suggested to be validated for internationally collaborated projects. A further study on synergy will also be run to measure the strength of SME collaboration synergy in the supply chain of power games.

REFERENCES

[1] D. Yuhas, ‘Sharing the Wealth (of Knowledge): Cumulative Cultural Development May Be Exclusively Human’, *Scientific American*, March 14, 2012, www. Scientificamerican.com, accessed: April, 11,2012.

[2] L. Tampieri, ‘The Governance of Synergies and Conflicts in Project Management: The Case of IPA Project RecoURB’, *Journal of the Knowledge Economy*, 1868-7873, (2011).

[3] A. Toppila, L. Juuso and A. Salo ‘A Resource Allocation Model for R&D Investments: A Case Study in Telecommunication Standardization’, *International Series in Operations Research & Management Science*, 162, 241-258 (2011).

[4] T.K. Das and B.S. Teng, 'Partner Analysis and Alliance Performance', *Scandinavian Journal of Management*, 19(3), 279-308, (2003).

[5] E. Gardet.and C. Mothe, 'SME Dependence and Coordination in Innovation Networks', *Journal of Small Business and Enterprise Development*, 19(2), 263-280, (2012).

[6] N. Pangarkar, 'Determinants of Alliance Duration in Uncertain Environments: The Case of the Biotechnology Sector', *Long Range Planning*,36(3), 269-284 (2003) .

[7] A. Wong, D. Tjosvold, P. Zhang, 'Developing Relationships in Strategic Alliances: Commitment to Quality and Cooperative Interdependence', *Industrial Marketing Management*, 34(7), 722-731, (2005).

[8] W.W. McCutchen Jr. and P.M. Swamidass, 'Motivations for Strategic Alliances In The Pharmaceutical/Biotech Industry: Some New Findings', *The Journal of High Technology Management Research*, 15(2), 197-214, (2004).

[9] Y.A. Hajidimitrou and A.C. Georgiou, 'A Goal Programming Model for Partner Selection Decisions in International Joint Ventures', European Journal of Operations Research, 3(1), 649-662, (2002).

[10] J.J. Huang, C.Y. Chen, H.H. Liu, G.H. Tzeng, 'A Multiobjective Programming Model For Partner Selection-Perspectives Of Objective Synergies And Resource Allocations', *Expert Systems with Applications*, 37(5), 3530-3536, (2010).

[11] A.P. Engelbrecht, *Fundamentals of Computational Swarm Intelligence*, Wiley Interscience, 2005, New York.

[12] A. Léon-Javier, N. Cruz-Cortez, M.A. Moreno-Armendariz and S. Orantes-Jiménez, “Finding Minimal Addition Chains with a Particle Swarm Optimization Algorithm”, *Lecture Notes in Computer Science, 2009, Volume 5845, MICAI 2009: Advances in Artificial Intelligence*, 2009, pp. 680-691.

[13] A. Altay *A Collective Intelligence Model for Assessing Collaborative Innovation Power Including Risks*, PhD. Dissertation, Istanbul Technical University, Industrial Engineering (2012).

[14] M.Levy and P. ‘SME Flexibility and the Role of Information Systems’, *Small Business Economics*, 11(8), 183-196 (1998).

[15] I.A. Khan ‘Knowledge Groups: A Model for Creating Synergy Across the Public Sector’ *Public Organization Review* (10)139–152 (2010).

[16] G. Chen and D. Tjosvold ‘Organizational values and procedures as antecedents for goal interdependence and collaborative effectiveness’ *Asia Pacific Journal of Management* (25:1) 93-112 (2007).

[17] S. Knoll, *Cross-Business Strategies*, Dissertation-Universitat St Gallen, Germany, Springer, 2008.

[18] E. Cefis and O. Mrsili, ‘Going, going, gone. Exit forms and the innovative capabilities of firms’, *Research Policy* (41:5) 795-807 (2012).

[19] M. Meuleman and W.De Maesenerie, ‘Do R&D subsidies affect SMEs’ access to external financing?’, *Research Policy*, (41:3) 580-591(2012).

[20] Y.C. Chang and C.J. Hsu ‘Ally or Merge – Airline Strategies After the Relaxation of Ownership Rules’, *Journal of the Eastern Asia Society for Transportation Studies*, (5) 545-556 (2005).

[21] Y. Lee, J. Shin and Y. Park, ‘The changing pattern of SME’s innovativeness through business model globalization’, *Technological Forecasting and Social Change*, (79:5) 832-842, 2012.

[22] S. Radas and L. Božić ‘The antecedents of SME innovativeness in an emerging transition economy’ *Technovation*, (29:6–7) 38-450, 2009.

[23] M. Brede, F. Boschetti and D. McDonald, ‘Managing Renewable Resources via Collective Intelligence’ *ModSim07*, Christchurch, New Zealand, December 2007.

[24] D. Brunker , *Collaboration and Other Factors Influencing Innovation Novelty in Australian Business*, Commonwealth of Australia, Department of Industry, Tourism and Resources (DITR) 2006.

[25] D. Twardy, *Partner Selection: A Source of Alliance Success, Research Projec*t, Eindhoven University of Technology and Zuyd University of Applied Science (2009).

[26] C Margoluis, *Healthy Relationships: Examining Alliances Within Population – Health – Environment Projects*, World Wildlife Fund Report (2008).

[27] A. Rai, S. Borah and A. Ramaprasad, ‘Critical Success Factors for Strategic Alliances in the Information Technology Industry: An Empirical Study’, *Decision Sciences*, (27:1) 141-155 (1996).

[28] J.F. Ding, ‘Partner Selection of Strategic Alliance for a Liner Shipping Company Using Extent Analysis Method of Fuzzy AHP’, *Journal of Marine Science and Technology*, (17:2), 97-105 (2009).

[29] B. Ramaseshan and P.C. Loo, ‘Factors Affecting a Partner’s Perceived Effective of Strategic Business Alliance: Some Singaporean Evidence’, *International Business Review*, (7:4), 443-458 (1998).

[30] S.H. Chen, H.T. Lee and Y.F. Wu, ‘Applying ANP Approach to Partner Selection for Strategic Alliance’, *Management Decision*, (46:3) 449-465 (2008).

[31] B. Gomes-Cassares, *Alliance Strategy: Fundamentals for Success*, Management Roundtable Workshop (2003).

[32] J.C. Linder, S. Perkins and P. Dover, *Drug Industry Alliances: In Search of Strategy*, Accenture White Paper (2004).

[33] Sherwood, A., Saxton, Inkpen, A. and Holzinger, I. ‘An Empirical Examination of the Relationships between Alliance Trust’, *10th International Conference on Reputation, Image, Identity and Competitiveness*, New York, (2006).

[34] L. Eden, *Friends, Acquaintances or Stranger? Partner Selection in R&D Alliances*, Texas A&M University, Draft Working Paper (2007).

[35] M.C. Tresch, V.C.K. Cheung and A. d’Avella, ‘Matrix Factorization Algorithms for the Identification of Muscle Synergies: Evaluation on Simulated and Experimental Data Sets’, *Journal of Neurophysiology*, (95) 2199-2212 (2005).

[36] W. Nelson, *Applied Life Data Analysis*, Wiley Series in Probability and Statistics, New York, (2004).

[37] S. Ross, *A First Course in Probability*, Prentice Hall, New York (2006).

[38] H. benRejeb, L. Morel- Guimarães, V. Boly and N’D. G. Assiélou ‘Measuring innovation best practices: Improvement of an innovation index integrating threshold and synergy effects’ *Technovation*, (28:12) 838-854 (2008).

[39] A. Altay and G. Kayakutlu *Animal Food Chain Based Particle Swarm Optimization*, World Congress on Engineering 2011, London, UK, 6-8 July 2011, Proceedings Edited by S. I. Ao, Len Gelman, David WL Hukins, Andrew Hunter, A. M. Korsunsky, II, 1094-1099 (2011).

[40] M.Q. Sulton and E.N. Anderson, Introduction to Cultural Ecology, Altamira Press, (2004).

[41] Press, (2011).

[42] B. Jarboui, M. Cheikh and P. Siarry, A. Rebai, ‘Combinatorial Particle Swarm Optimization (CPSO) for Partitional Clustering Problem’, *Applied Mathematics and Computation*, (192:2) 337-345 (2007).

[43] R. Gras, D. Devaurs, A. Wozniak and A. Aspinall, ‘An Individual-Based Evolving Predator-Prey Ecosystem Simulation Using a Fuzzy Cognitive Map as the Behavior Model’, *Artificial Life*, (15:4) 423-463 (2009).