



**INFORMATION SCIENCE  
&  
COMPUTING**

---

International Book Series

Number 14

---

**New Trends  
in  
Intelligent Technologies**

---

Supplement to  
International Journal "Information Technologies and Knowledge" Volume 3 / 2009

---

**ITHEA  
SOFIA, 2009**

Luis Fernando de Mingo Lopez, Juan Castellanos,  
Krassimir Markov, Krassimira Ivanova, Iliya Mitov (ed.)

**New Trends in Intelligent Technologies**

International Book Series "INFORMATION SCIENCE & COMPUTING", Number 14

Supplement to the International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Volume 3 / 2009

Institute of Information Theories and Applications FOI ITHEA

Sofia, Bulgaria, 2009

This issue contains a collection of papers that concern the problems of intelligent technologies.

Papers in this issue are selected from the International Conference i.TECH-2 2009, Madrid, Spain, a part of the Joint International Events of Informatics "ITA 2009", Autumn Session.

International Book Series "INFORMATION SCIENCE & COMPUTING", Number 14  
Supplement to the International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Volume 3, 2009

Edited by Institute of Information Theories and Applications FOI ITHEA, Bulgaria,  
and Universidad Politecnica de Madrid, Spain,  
in collaboration with Institute of Mathematics and Informatics, BAS, Bulgaria,  
and Institute of Information Technologies, BAS, Bulgaria.

Publisher: Institute of Information Theories and Applications FOI ITHEA, Sofia, 1000, P.O.B. 775, Bulgaria  
Издател: Институт по информационни теории и приложения ФОИ ИТЕА, София, 1000, п.к. 775, България  
[www.ithea.org](http://www.ithea.org), [www.foibg.com](http://www.foibg.com), e-mail: [info@foibg.com](mailto:info@foibg.com)

General Sponsor: Consortium FOI Bulgaria ([www.foibg.com](http://www.foibg.com)).

Printed in Bulgaria

Copyright © 2009 All rights reserved

© 2009 Institute of Information Theories and Applications FOI ITHEA - Publisher

© 2009 Luis Fernando de Mingo Lopez, Juan Castellanos,  
Krassimir Markov, Krassimira Ivanova, Iliya Mitov – Editors

© 2009 For all authors in the issue.

ISSN 1313-0455 (printed)

ISSN 1313-048X (online)

ISSN 1313-0501 (CD/DVD)

---

## PREFACE

The scope of the International Book Series "Information Science and Computing" (IBS ISC) covers the area of Informatics and Computer Science. It is aimed to support growing collaboration between scientists from all over the world. IBS ISC is official publisher of the works of the members of the ITHEA International Scientific Society.

The official languages of the IBS ISC are English and Russian.

IBS ISC welcomes scientific papers and books connected with any information theory or its application.

IBS ISC rules for preparing the manuscripts are compulsory.

The rules for the papers and books for IBS ISC are given on [www.foibg.com/ibsisc](http://www.foibg.com/ibsisc) .

The camera-ready copies of the papers should be received by ITHEA Submission System <http://ita.ithea.org> .

The camera-ready copies of the books should be received by e-mail: [info@foibg.com](mailto:info@foibg.com) .

Responsibility for papers and books published in IBS ISC belongs to authors.

This issue contains a collection of papers that concern the problems of intelligent technologies. Papers are peer reviewed and are selected from the International Conference i.TECH-2, Madrid, Spain, a part of the Joint International Events of Informatics "ITA 2009" – autumn session.

ITA 2009 has been organized by

ITHEA International Scientific Society

in collaboration with:

- Universidad Politecnica de Madrid (Spain)
- Institute of Information Theories and Applications FOI ITHEA
- International Journal "Information Theories and Applications"
- International Journal "Information Technologies and Knowledge"
- Association of Developers and Users of Intelligent Systems (Ukraine)
- Association for Development of the Information Society (Bulgaria)
- V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine
- Institute of Mathematics and Informatics, BAS (Bulgaria)
- Institute of Information Technologies, BAS (Bulgaria)
- Dorodnicyn Computing Centre of the Russian Academy of Sciences
- Institute of Mathematics of SD RAN (Russia)
- Taras Shevchenko National University of Kiev (Ukraine)
- BenGurion University (Israel)
- University of Calgary (Canada)
- University of Hasselt (Belgium)
- Kharkiv National University of Radio Electronics (Ukraine)
- Rzeszow University of Technology (Poland)
- Astrakhan State Technical University (Russia)
- Varna Free University "Chernorizets Hrabar" (Bulgaria)
- National Laboratory of Computer Virology, BAS (Bulgaria)
- Uzhgorod National University (Ukraine)

The main ITA 2009 events were:

KDS	XVth International Conference "Knowledge - Dialogue – Solution"
i.Tech	Seventh International Conference "Information Research and Applications"
MeL	Fourth International Conference "Modern (e-) Learning"
INFOS	Second International Conference "Intelligent Information and Engineering Systems"
CFDM	International Conference "Classification, Forecasting, Data Mining"
GIT	Seventh International Workshop on General Information Theory
ISSI	Third International Summer School on Informatics

More information about ITA 2009 International Conferences is given at the [www.ithea.org](http://www.ithea.org).

The great success of ITHEA International Journals, International Book Series and International Conferences belongs to the whole of the ITHEA International Scientific Society.

We express our thanks to all authors, editors and collaborators who had developed and supported the International Book Series "Information Science and Computing".

General Sponsor of IBS ISC is the Consortium FOI Bulgaria ([www.foibg.com](http://www.foibg.com)).

*Madrid-Sofia, September 2009*

*L. F. de Mingo Lopez, J.Castellanos, Kr. Markov, Kr.Ivanova, I.Mitov*

---

**TABLE OF CONTENTS**

<i>Preface</i> .....	3
<i>Table of Contents</i> .....	5
<i>Index of Authors</i> .....	7
<b>Simulation of DNA Cutting</b>	
<i>Francisco José Cisneros, Andrés de la Peña, Cristina Piqueras, Paula Cordero, Juan Castellanos</i> .....	9
<b>P-Systems: Study of Randomness when Applying Evolution Rules</b>	
<i>Alberto Arteta, Luis Fernández, Fernando Arroyo</i> .....	15
<b>Modeling Language of Multi-agent Systems = Programming Template</b>	
<i>Rubén Álvarez-González, Miguel Angel Díaz Martínez</i> .....	24
<b>Comparison of Discretization Methods for Preprocessing Data for Pyramidal Growing Network Classification Method</b>	
<i>Iliia Mitov, Krassimira Ivanova, Krassimir Markov, Vitalii Velychko, Peter Stanchev, Koen Vanhoof</i> .....	31
<b>Multilanguage Opera Subtitling Exchange between Production and Broadcaster Companies</b>	
<i>Jesús Martínez Barbero, Manuel Bollain Pérez</i> .....	40
<b>Performance Analysis of Call Admission Control for Streaming Traffic with Activity Detection Function</b>	
<i>Kiril Kassev, Yakim Mihov, Boris Tsankov</i> .....	47
<b>Analysis of Malicious Attacks Accomplished in Real and Virtual Environment</b>	
<i>Dimitrina Polimirova, Eugene Nickolov</i> .....	53
<b>Digital Objects – Storage, Delivery and Reuse</b>	
<i>Juliana Peneva, Stanislav Ivanov, Filip Andonov, Nikolay Dokev</i> .....	61
<b>Multi-modal Emotion Recognition – More "Cognitive" Machines</b>	
<i>Velina Slavova, Hichem Sahli, Werner Verhelst</i> .....	70
<b>Prognostication of Efficiency of Medical and Prophylactic Measures at Different Homeostasis Violation of Human Organism by Markov Processes Theory</b>	
<i>Boris Samura, Anatoly Povoroznuk, Olga Kozina, Elena Visotskaja, Elena Chernykh, Nikolay Shukin, Andrei Porvan</i> .....	79

## Indirect Spatial Data Extraction from Web Documents

*Dimitar Blagoev, George Totkov, Milena Staneva, Krassimira Ivanova, Krassimir Markov, Peter Stanchev.... 89*

## Adaptation for Assimilation: the Role of Adaptable m-Learning Services in the Modern Educational Paradigm

*Damien Meere, Ivan Ganchev, Stanimir Stojanov, Máirtín O’Dróma ..... 101*

## EulerPathSolver: a New Application for Fleury’s Algorithm Simulation

*Gloria Sánchez–Torrubia, Carmen Torres–Blanc, Leila Navascués-Galante ..... 111*

## Automatic Metadata Generation for Specification of e-Documents – the METASPEED Project

*Juliana Peneva, George Totkov, Peter Stanchev, Elena Shoikova ..... 118**About ITA 2010: Joint International Scientific Events on Informatics ..... 127*

## INDEX OF AUTHORS

Ruben Alvarez-Gonzalez	24	Eugene Nickolov	53
Filip Andonov	61	Mairtin O'Droma	101
Fernando Arroyo	15	Andres de la Pena	9
Alberto Arteta	15	Juliana Peneva	61, 118
Dimitar Blagoev	89	Cristina Piqueras	9
Manuel Bollain Perez	40	Dimitrina Polimirova	53
Juan Castellanos	9	Andrei Porvan	79
Elena Chernykh	79	Anatoly Povoroznuk	79
Francisco Jose Cisneros	9	Hichem Sahli	70
Paula Cordero	9	Boris Samura	79
Miguel Angel Diaz	24	Gloria Sanchez-Torrubia	111
Nikolay Dokev	61	Elena Shoikova	118
Luis Fernandez	15	Nikolay Shukin	79
Ivan Ganchev	101	Velina Slavova	70
Stanislav Ivanov	61	Peter Stanchev	31, 89, 118
Krassimira Ivanova	31, 89	Milena Staneva	89
Kiril Kashev	47	Stanimir Stojanov	101
Olga Kozina	79	Carmen Torres-Blanc	111
Krassimir Markov	31, 89	George Totkov	89, 118
Jesus Martinez Barbero	40	Boris Tsankov	47
Damien Meere	101	Koen Vanhoof	31
Yakim Mihov	47	Vitalii Velychko	31
Ilia Mitov	31	Werner Verhelst	70
Leila Navascues-Galante	111	Elena Visotskaja	79





---

## SIMULATION OF DNA CUTTING

Francisco José Cisneros, Andrés de la Peña, Cristina Piqueras,  
Paula Cordero, Juan Castellanos

*Abstract:* The simulation of the main molecular operations used in DNA Computing can lead the researchers to develop complex algorithms and methods without the need of working with real DNA strands in-vitro. The purpose of this paper is to present a computer program which simulates a cutting process over DNA molecules which is an essential operation for the DNA computation. This simulation represents a useful tool for a virtual laboratory which is oriented to DNA computations. The results given by the software can show the behavior of a DNA cutting under certain set of restrictive enzymes to carry out the operation in-vitro efficiently.

*Keywords:* DNA Computing, DNA Simulation, Software Simulation, DNA operations, Bioinformatics.

*ACM Classification Keywords:* I.6. Simulation and Modelling, B.7.1 Advanced Technologies, J.3 Biology and Genetics

---

### Introduction

---

DNA Computing is an impressive computer paradigm based on the work made by Leonard M. Adleman [Adleman, 1994], where the first implementation of a computer based on DNA operations solved a hard combinatorial problem using deoxyribonucleic acid molecules. He was able to solve an NP-complete problem using DNA molecules and biological operations. This represented an approach to a massive parallel paradigm.

Molecular computing consists of representing the information of the problem with organic molecules [J.Castellanos, 1998] and to make them react within a test tube in order to solve a problem. The fundamental characteristics of this type of computations are, mainly, the massive parallelism of DNA strands and the Watson-Crick complementarity. The speed of calculation, the small consumption of energy and the big amount of information which DNA strands are able to store are the best advantages that DNA computing has. Nevertheless one of the problems is the massive calculation space needed, which limits the size of the problems.

The nucleic acids are linear polymers in which the repetitive unit is the nucleotide. Each nucleotide is formed by a pentose (the ribose or the deoxyribose), a nitrogenous base (purin or pyrimidin) and a phosphoric acid. The union of the pentose with a base constitutes the nucleoside. The union of this last structure with the phosphoric acid gives us the nucleotide. The union of the nucleotides gives us the polynucleotide. The nitrogenous bases that form each DNA molecule are Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). Those which form each RNA molecule are Adenine (A), Guanine (G), Cytosine (C) and Uracil (U). Double stranded molecules are formed by two strands twisted in a helix. The Adenine of a helix matches the Thymine of the complementary helix by creating two hydrogenate bridges. Also, the Guanine of a helix matches the Cytosine of the complementary three hydrogenate bridges. Therefore, the bases of one strand are united by hydrogenate bridges to the bases of the other strand, forming the base pairs AT and GC.

It is very important to determine which biologic operations could be used for the manipulation of DNA strands. In order to distinguish between the common mathematical operations and the biological procedures which are applied on DNA strands, it is used the term bio-operations to talk about the last ones. Some of the bio-operations that facilitate the manipulation of DNA are the measure of DNA strands, the DNA cutting, the lengthening and shortening of DNA strands, the separation and fusion of DNA sequences (denaturalization and renaturalization) or the ordination by length or electrophoresis among others [2].

In this article it is explained the development of a software that simulates successfully the process of cutting over DNA molecules. The aim of it is to incorporate this cutting tool to a virtual laboratory in which all the operations explained above are implemented. This environment help us to prove how molecules would react to the codifications we develop in-info so that the steps needed in a real laboratory are reduce substantially.

## DNA cutting

The DNA cutting is one of the most basic operation in computing with DNA, this is so because allow to manipulate the DNA strands in specific points. The data problem are represented through nucleotides sequences, an, at the same time, this sequences, which represent the problem atomic data, are linked in bigger sequences, representing data sets or lists which normally are associated to possible solutions. In order to carry out the manipulation, such as atomic elements extraction, elimination, combination or addition to the possible solutions, generally is necessary to apply the cutting operation. This cutting operation is carried out applying restriction enzymes, which are in charge of making the operation in a parallel massive manner [Kobayashi, 2001].

Enzymes are biomolecules that catalyze chemical reactions. Restriction enzymes (or restriction endonuclease), found in bacteria and archaea, are enzymes that cuts double-stranded or single stranded DNA at specific recognition nucleotide sequences known as restriction sites [Roberts, 2007]. There are three types of restriction enzyme. Such that are included in type I are characteristic of two different strains of *E. coli*. These enzymes cut at a site that differs, and is some distance (at least 1000 bp) away, from their recognition site. The recognition site is asymmetrical and is composed of two portions – one containing 3-4 nucleotides, and another containing 4-5 nucleotides – separated by a spacer of about 6-8 nucleotides. Type II restriction enzymes [Roberts, 2005] are composed of only one subunit, their recognition sites are usually undivided and palindromic and 4-8 nucleotides in length, and they recognize and cleave DNA at the same site. Type III restriction enzymes recognize two separate non-palindromic sequences that are inversely oriented. They cut DNA about 20-30 base pairs after the recognition site.

Type II enzymes are the most commonly available and used restriction enzymes. They are specially good in computing with DNA due to the atomic data are codify frequently for space reasons, with nucleotides sequences the shortest as possible, often smaller than the distance between the cutting point and the recognition site.

Examples of restriction enzymes include [Roberts, 1980]:

Enzyme	Source	Recognition Sequence	Cut
EcoRI	<i>Escherichia coli</i>	5'GAATTC 3'CTTAAG	5'---G AATTC---3' 3'---CTTAA G---5'
EcoRII	<i>Escherichia coli</i>	5'CCWGG 3'GGWCC	5'--- CCWGG---3' 3'---GGWCC ---5'
Smal*	<i>Serratia marcescens</i>	5'CCCGGG 3'GGGCCC	5'---CCC GGG---3' 3'---GGG CCC---5'

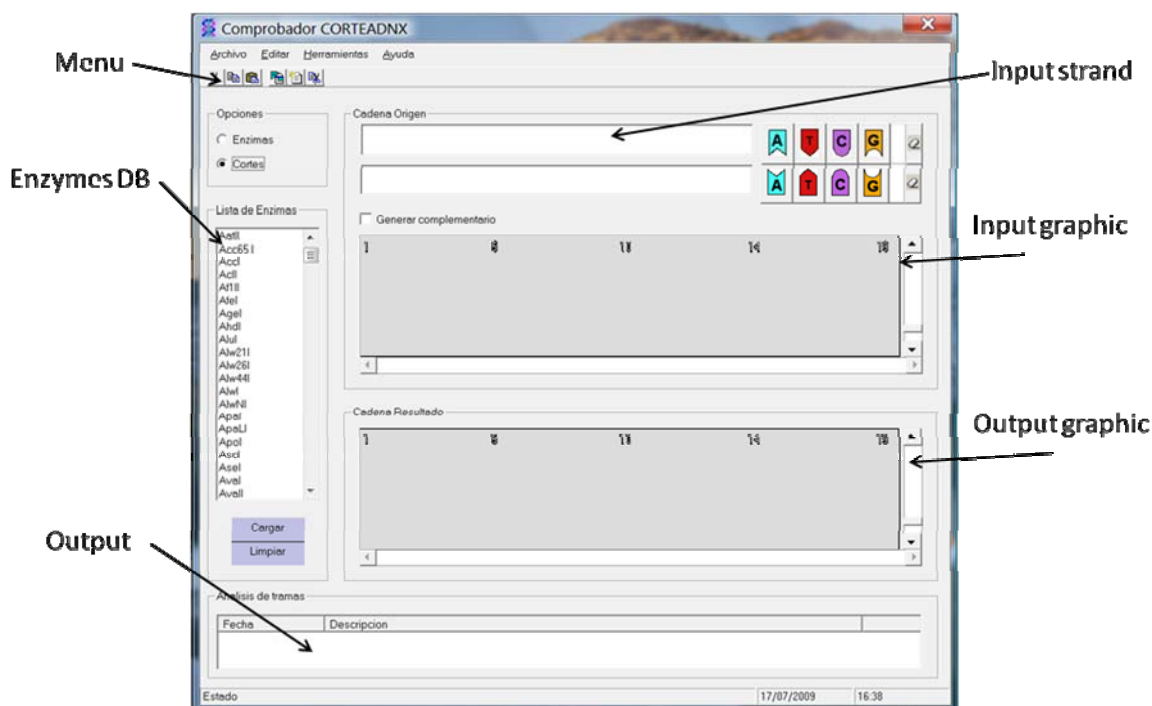
The number of commercial available restriction enzymes is constantly increasing [Roberts, 2003], fact that allow to improve the computation algorithm, since this ones are often limited, especially due to the relatively reduce amount of enzymes with known cut.

## DNA cutting simulation

The software of simulation realized allows us establish all recognition sites of DNA strand, and how remain the cuts and the resultants sub chains. The simulator development has being dealt in four parts, the interface, an enzymes database, the cutting simulator and the obtained results viewer, this tow last are ActiveX objects, allowing its use in others simulators, in such a way is possible to combine this ones in a common way. The entire application is written in Visual BASIC 6.

The data access is carried out through and interfaces ODBC, which allow establishing the connection with a wide amount of existing database and in a constant updating process. This is an especially important factor given the permanent emerging of new enzymes allowing, not only enriching languages, but also, in some cases, they contribute with the capacity of carry out operations that improve the algorithm performance. It is attached a database with the 280 more frequently used enzymes.

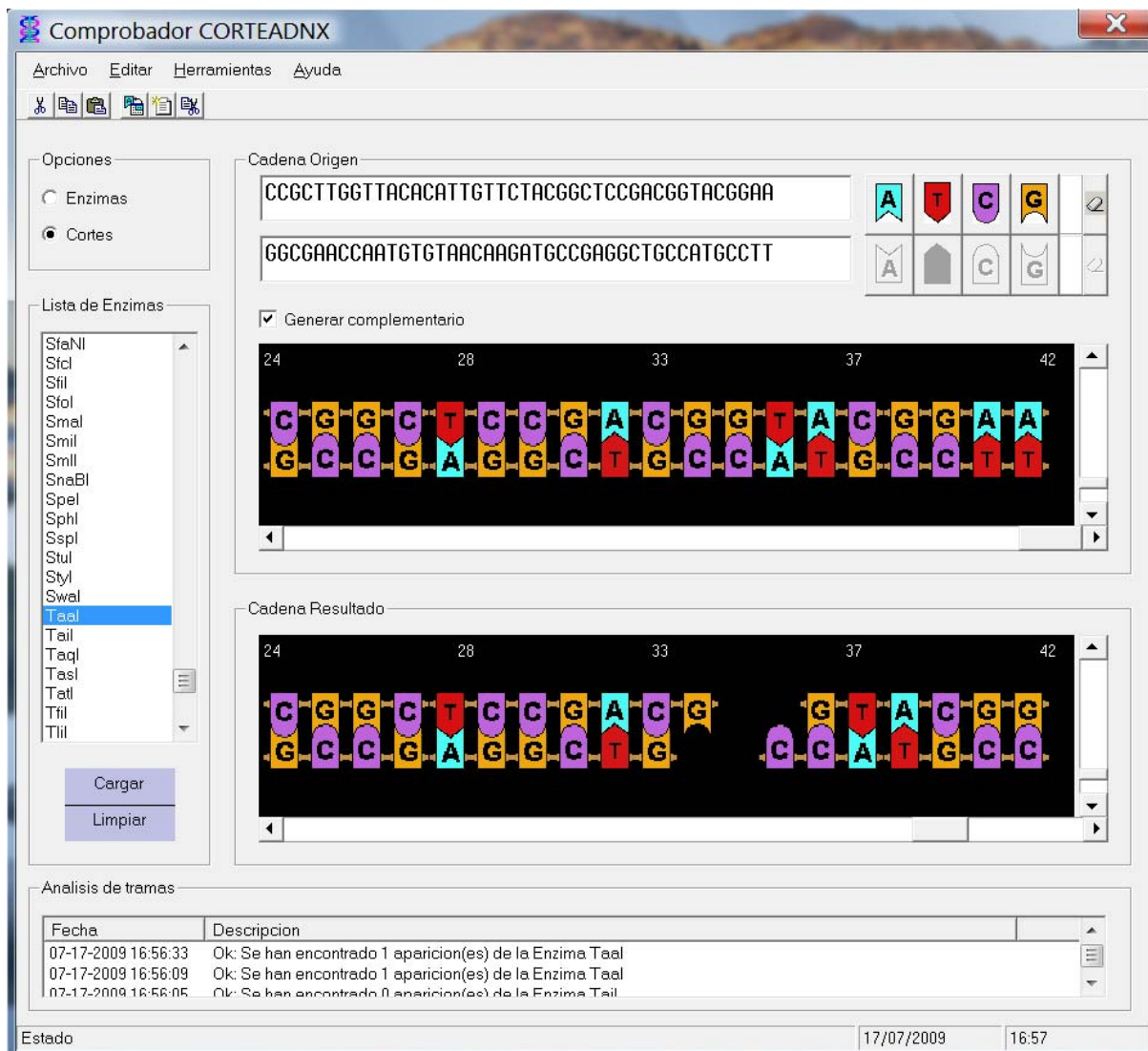
The interface is easy and intuitive, in such a way it could be used without the need of a previous formative course; the X Figure shows the different simulator aspects.



The simulator allow to visualize graphically the application result of an enzyme over a given chain, or, in other way, it is possible to generate a result report of applying every enzymes available over the strand, identifying which find out its recognition site.

## Example

Bellow its is presented an example where it is detected the entire amount of enzymes which have a recognition site, content in the chain introduced in order to know the enzymes which could cut the chain. It is a typical example, showing when the anti complementary and stable languages are designed, which need to count with ruptur points, assuring in such a way that no one enzyme will produce a cut in an improper site.



As we can see in the results and reviewing the report that gives us the simulator, we found that the chain can be cut with five enzymes (BspLI, Csp6I, NlaIV, RsaI, Taal). The chosen screenshot shows the cut made by Taal enzyme as a checking.

## Conclusion

The simulation model here presented allows us to carry out experiments with DNA in a simple way using only our computer. By using it we are able to understand the behavior of DNA molecules in a cutting process without the costs, the time and the space needed in a laboratory. That is why it exist the need of making these experiments easier by simplify the main bio-operations and bio-molecular processes over DNA molecules. It is very important and useful to develop this kind of software simulators so that researchers can create more complicated algorithms for DNA computations without the need of testing in a DNA pool every single operation.

This software try to give a simple view of the behavior of DNA molecules when it is applied a certain cutting enzyme. This shows the exact cutting point and the final shape of each resulting strands, and also is able to recognize the entire amount of enzymes that could cut the chain. The use of this simulator can give a result of a cutting process of the DNA molecules introduced in a faster way than in-vitro.

---

**Bibliography**

---

- [Adleman, 1994] Leonard M. Adleman. Molecular Computation of Solutions to Combinatorial Problems. *Science* (journal) 266 (11): 1021-1024. 1994.
- [Adleman, 1998] Leonard M. Adleman. Computing with DNA. *Scientific American* 279: 54-61. 1998
- [Lipton, 1995] Richard J. Lipton. Using DNA to solve NP-Complete Problems. *Science*, 268:542-545. April 1995
- [J.Castellanos, 1998] J.Castellanos, S.Leiva, J.Rodrigo, A. Rodríguez Patón. Molecular computation for genetic algorithms. First International Conference, RSCTC'98.
- [Shannon, 1949] C.E.Shannon. The Mathematical theory of communication. In: *The Mathematical Theory of Communication*. Ed. C.E.Shannon and W.Weaver. University of Illinois Press, Urbana, 1949.
- [Roberts, 1980] Roberts RJ. [Restriction and modification enzymes and their recognition sequences](#). *Nucleic Acids Res.* 8 (1): r63-r80. [doi:10.1093/nar/8.1.197-d](#). [PMID 6243774](#).
- [Roberts, 1976] Roberts RJ. Restriction endonucleases. *CRC Crit. Rev. Biochem.* 4 (2): 123-64. [doi:10.3109/10409237609105456](#). [PMID 795607](#).
- [Kessler, 1990] Kessler C, Manta V. Specificity of restriction endonucleases and DNA modification methyltransferases a review (Edition 3). *Gene* 92 (1-2): 1-248. [doi:10.1016/0378-1119\(90\)90486-B](#). [PMID 2172084](#).
- [Pingoud, 1993] Pingoud A, Alves J, Geiger R. Chapter 8: Restriction Enzymes. in Burrell, Michael. *Enzymes of Molecular Biology. Methods of Molecular Biology*. 16. Totowa, NJ: Humana Press. pp. 107-200. [ISBN 0-89603-234-5](#).
- [Kobayashi, 2001] Kobayashi I. [Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution](#). *Nucleic Acids Res.* 29 (18): 3742-56. [doi:10.1093/nar/29.18.3742](#). [PMID 11557807](#).
- [Roberts, 2005] Roberts RJ (April 2005). [How restriction enzymes became the workhorses of molecular biology](#). *Proc. Natl. Acad. Sci. U.S.A.* 102 (17): 5905-8. [doi:10.1073/pnas.0500923102](#). [PMID 15840723](#). [PMC: 1087929](#). <http://www.pnas.org/cgi/pmidlookup?view=long&pmid=15840723>.
- [Roberts, 2007] Roberts RJ, Vincze T, Posfai J, Macelis D. (2007). [REBASE--enzymes and genes for DNA restriction and modification](#). *Nucleic Acids Res* 35 (Database issue): D269-70. [doi:10.1093/nar/gkl891](#). [PMID 17202163](#). <http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=17202163>.
- [Adrienne, 2001] Adrienne Massey; Helen Kreuzer (2001). *Recombinant DNA and Biotechnology: A Guide for Students*. Washington, D.C: ASM Press. [ISBN 1-55581-176-0](#).
- [Pingoud, 2001] Pingoud A, Jeltsch A (September 2001). [Structure and function of type II restriction endonucleases](#). *Nucleic Acids Res.* 29 (18): 3705-27. [doi:10.1093/nar/29.18.3705](#). [PMID 11557805](#).
- [Roberts, 2003] Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, Blumenthal RM, Degtyarev SKh, Dryden DT, Dybvig K, Firman K, Gromova ES, Gumpert RI, Halford SE, Hattman S, Heitman J, Hornby DP, Janulaitis A, Jeltsch A, Josephsen J, Kiss A, Klaenhammer TR, Kobayashi I, Kong H, Krüger DH, Lacks S, Marinus MG, Miyahara M, Morgan RD, Murray NE, Nagaraja V, Piekarowicz A, Pingoud A, Raleigh E, Rao DN, Reich N, Repin VE, Selker EU, Shaw PC, Stein DC, Stoddard BL, Szybalski W, Trautner TA, Van Etten JL, Vitor JM, Wilson GG, Xu SY. [A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes](#). *Nucleic Acids Res.* 31 (7): 1805-12. [doi:10.1093/nar/gkg274](#). [PMID 12654995](#).
- [Geerlof, 2008] Geerlof A. [Cloning using restriction enzymes](#). European Molecular Biology Laboratory - Hamburg. [http://www.embl-hamburg.de/~geerlof/webPP/genetoprotein/cloning\\_strategy/clo\\_rest-enzymes.html](http://www.embl-hamburg.de/~geerlof/webPP/genetoprotein/cloning_strategy/clo_rest-enzymes.html).
- [Wolff, 2008] Wolff JN, Gemell NJ. Combining allele-specific fluorescent probes and restriction assay in real-time PCR to achieve SNP scoring beyond allele ratios of 1:1000. *BioTechniques* 44 (2): 193-4, 196, 199. [doi:10.2144/000112719](#). [PMID 18330346](#).
- [Zhang, 2005] Zhang R, Zhu Z, Zhu H, Nguyen T, Yao F, Xia K, Liang D, Liu C. ["SNP Cutter: a comprehensive tool for SNP PCR-RFLP assay design"](#). *Nucleic Acids Res.* 33 (Web Server issue): W489-92. [doi:10.1093/nar/gki358](#). [PMID 15980518](#).

**Authors' Information**

---

*Francisco J. Cisneros* – Natural Computing Group, Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain, e-mail: [kikocisneros@gmail.com](mailto:kikocisneros@gmail.com)

*Andrés de la Peña* – Natural Computing Group, Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain, e-mail: [andres.de.la.pena@gmail.com](mailto:andres.de.la.pena@gmail.com)

*Cristina Piqueras* – Natural Computing Group, Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain

*Paula Cordero* – Natural Computing Group, Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain, e-mail: [paula.cormo@gmail.com](mailto:paula.cormo@gmail.com)

*Juan Castellanos* – Natural Computing Group, Artificial Intelligence Department, Facultad de Informática, Universidad Politécnica de Madrid, Boadilla del Monte, 28660. Madrid, Spain, e-mail: [jcastellanos@fi.upm.es](mailto:jcastellanos@fi.upm.es)

---

## P-SYSTEMS: STUDY OF RANDOMNESS WHEN APPLYING EVOLUTION RULES

Alberto Arteta, Luis Fernández, Fernando Arroyo

*Abstract:* Membrane computing is a recent area that belongs to natural computing. This field works on computational models based on nature's behavior to process the information. Recently, numerous models have been developed and implemented with this purpose. P-systems are the structures which have been defined, developed and implemented to simulate the behavior and the evolution of membrane systems which we find in nature. What we analyze in this paper is the power of the tools we currently have to simulate the randomness we find in nature. The main problem we face here, is trying to simulate non deterministic events by using deterministic tools. The goal we want to achieve is to propose an optimal method when simulating non deterministic processes. Talking about simulation of non deterministic method makes no sense when using deterministic tools; however we can get closer to the idea of non determinism by using more powerful randomness generators.

*Keywords:* P-systems, evolution rules application, non-determinism simulation, randomness in p-systems

---

### Introduction

---

Natural computing is a new field within computer science which develops new computational models. These computational models can be divided into three major areas:

- Neural networks.
- Genetic Algorithms
- Biomolecular computation.

Membrane computing is included in biomolecular computation. Within the field of membrane computing a new logical computational device appears: The P-system. These P-systems are able to simulate the behavior of the membranes on living cells. This behavior refers to the way membranes process information. (Absorbing nutrients, chemical reactions, dissolving, etc)

Membrane computing formally represents, through the use of P-systems, the processes that take place inside of the living cells. In terms of software systems, it is the process within a complex and distributed software. In parallel computational models, p-systems might be as important as the Turing machine is in sequential computational models.[Arroyo, 2001]

In this paper, we study the current methods to implement the idea of randomness. Most of the times the function rnd is used for that purpose. By doing that we state that an important part of inner quality on nature is missed. We will prove that such function has low quality on terms of randomness. When a p-system has a few evolution rules, this will not create any problem. However the entire simulation will degrade when the number of evolution rules increases. By proposing a new way of generating randomness we will get close to the idea of 'pure randomness' we find in nature and also we would be able to show a higher quality simulation.

In order to do this, we will take the following steps:

- Introduction to P-systems theory;
- Analysis of rules application process;
- Analysis of the Random Function
- Study of the current methods to implement non-determinism

- Proposal of a new method.
- Conclusions and further work.

## Introduction to P-systems Theory

In this section we will study into detail all of the theories related to the paradigm of the P-systems. A P-system is a computational model inspired by the way the living cells interact with each other through their membranes. The elements of the membranes are called objects. A region within a membrane can contain objects or other membranes. A p-system has an external membrane (also called skin membrane) and it also contains a hierarchical relation defined by the composition of the membranes. A multiset of objects is defined within a region (enclosed by a membrane). These multisets of objects show the number of objects existing within a region. Any object 'x' will be associated to a multiplicity which tells the number of times that 'x' is repeated in a region.

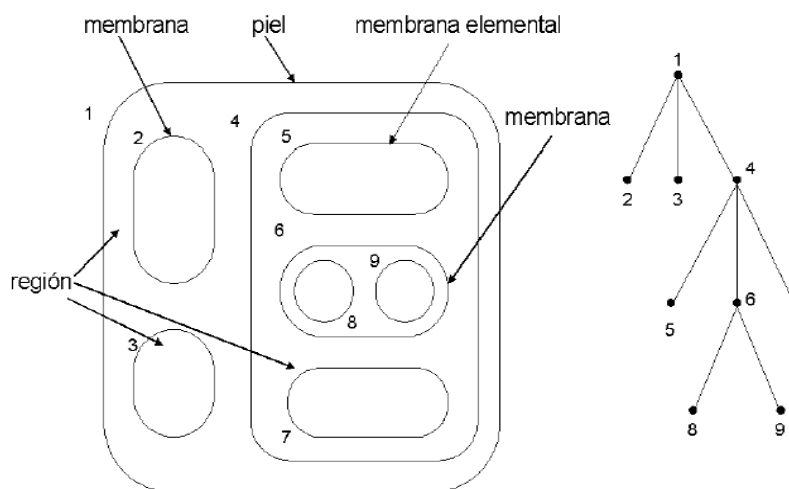


Fig. 1. The membrane's structure (left) represented in tree shape (right)

According to Păun's definition, a transition P System of degree  $n$ ,  $n > 1$  is a construct: [Păun 1998]

$$\Pi = (V, \mu, \omega_1, \dots, \omega_n, (R_1, \rho_1), \dots, (R_n, \rho_n), i_0)$$

where:

- $V$  is an alphabet; its elements are called objects;
- $\mu$  is a membrane structure of degree  $n$ , with the membranes and the regions labeled in a one-to-one manner with elements in a given set; in this section we always use the labels  $1, 2, \dots, n$ ;
- $\omega_i$   $1 \leq i \leq n$ , are strings from  $V^*$  representing multisets over  $V$  associated with the regions  $1, 2, \dots, n$  of  $\mu$
- $R_i$   $1 \leq i \leq n$ , are finite set of evolution rules over  $V$  associated with the regions  $1, 2, \dots, n$  of  $\mu$ ;  $\rho_i$  is a partial order over  $R_i$   $1 \leq i \leq n$ , specifying a priority relation among rules of  $R_i$ . An evolution rule is a pair  $(u, v)$  which we will usually write in the form  $u \rightarrow v$  where  $u$  is a string over  $V$  and  $v = v'$  or  $v = v' \delta$  where  $v'$  is a string over  $(V \times \{here, out\}) \cup (V \times \{in_j \mid 1 \leq j \leq n\})$ , and  $\delta$  is a special symbol not in  $V$ . The length of  $u$  is called the radius of the rule  $u \rightarrow v$
- $i_0$  is a number between 1 and  $n$  which specifies the output membrane of  $\Pi$



Let  $U$  be a finite and not an empty set of objects and  $N$  the set of natural numbers. A *multiset of objects* is defined as a mapping:

$$M : V \rightarrow N$$

$$a_i \rightarrow u_i$$

Where  $a_i$  is an object and  $u_i$  its multiplicity.

As it is well known, there are several representations for multisets of objects.

$$M = \{(a_1, u_1), (a_2, u_2), (a_3, u_3), \dots\} = a_1^{u_1} \cdot a_2^{u_2} \cdot a_n^{u_n} \dots$$

Evolution rule with objects in  $U$  and targets in  $T$  is defined by  $r = (m, c, \delta)$

where  $m \in M(V)$ ,  $c \in M(V \times T)$  and  $\delta \in \{\text{to dissolve, not to dissolve}\}$

From now on 'c' will be referred to as the consequent of the evolution rule 'r'

The set of evolution rules with objects in  $V$  and targets in  $T$  is represented by  $R(U, T)$ .

We represent a rule as:

$x \rightarrow y$  or  $x \rightarrow y\delta$  where  $x$  is a multiset of objects in  $M((V) \times \text{Tar})$  where  $\text{Tar} = \{\text{here, in, out}\}$  and  $y$  is the consequent of the rule. When  $\delta$  is equal to "dissolve", then the membrane will be dissolved. This means that objects from a region will be placed within the region which contains the dissolved region. Also, the set of evolution rules included on the dissolved region will disappear.

P-systems evolve, which makes it change upon time; therefore it is a dynamic system. Every time that there is a change on the p-system we will say that the P-system is in a new transition. The step from one transition to another one will be referred to as an evolutionary step, and the set of all evolutionary steps will be named computation. Processes within the p-system will be acting in a *massively parallel* and *non-deterministic* manner. (Similar to the way the living cells process and combine information).

We will say that the computation has been successful if:

1. The halt status is reached.
2. No more evolution rules can be applied.
3. Skin membrane still exists after the computation finishes.

---

## Analysis of Rules Application Process

---

In this paper we focus on the application of evolution rules. Every region of a p\_system contains a multiset of symbol-objects, which correspond to the chemicals swimming in a solution in a cell compartment; these chemicals are considered here as unstructured, that is why we describe them by symbols from a given alphabet.

The objects evolve by means of evolution rules, which are also localized, associated with the regions of the membrane structure. There are three main types of rules:[Păun 1998]

1. Multiset rewriting rules (one uses to call them, simply, evolution rules),
2. Communication rules,
3. Rules for handling membranes.

In this section we present the first type of rules. They correspond to the chemical reactions possible in the compartments of a cell, hence they are of the form  $u \rightarrow v$ , where  $u$  and  $v$  are multisets of objects. However, in order to make the compartments cooperate, we have to move objects across membranes, and to this aim we add

target indications to the objects produced by a rule as above (to the objects from multiset  $v$ ). These indications are: "*here, in, out*", with the meaning that an object having associated the indication *here* remains in the same region, one having associated the indication *in* goes immediately into a directly lower membrane, non-deterministically chosen, and *out* indicates that the object has to exit the membrane, thus becoming an element of the region surrounding it. An example of evolution rule is:

$$aab \rightarrow (a, \textit{here})(b, \textit{out})(c, \textit{here})(c, \textit{in})$$

(this is the first of the rules considered in Section 4, with target indications associated with the objects produced by rule application). After using this rule in a given region of a membrane structure, two copies of  $a$  and one  $b$  are consumed (removed from the multiset of that region), and one copy of  $a$ , one of  $b$ , and two of  $c$  are produced; the resulting copy of  $a$  remains in the same region, and the same happens with one copy of  $c$  (indications *here*), while the new copy of  $b$  exits the membrane, going to the surrounding region (indication *out*), and one of the new copies of  $c$  enters one of the child membranes, non-deterministically chosen. If no such child membrane exists, that is, the membrane with which the rule is associated is elementary, then the indication *in* cannot be followed, and the rule cannot be applied. In turn, if the rule is applied in the skin region, then  $b$  will exit into the environment of the system (and it is "lost" there, as it can never come back). In general, the indication *here* is not specified (an object without an explicit target indication is supposed to remain in the same region where the rule is applied).

A rule as above, with at least two objects in its left hand side, is said to be cooperative; a particular case is that of catalytic rules, of the form  $ca \rightarrow cv$ , where  $c$  is an object (called catalyst) which assists the object  $a$  to evolve into the multiset  $v$ ; rules of the form  $a \rightarrow v$ , where  $a$  is an object, are called non-cooperative.

The rules can also have the form  $u \rightarrow v \delta$ , where  $\delta$  denotes the action of membrane dissolving:

if the rule is applied, then the corresponding membrane disappears and its contents, object and membranes alike, are left free in the surrounding membrane; the rules of the dissolved membrane disappear at the same time with the membrane. The skin membrane is never dissolved.

The communication of objects through membranes reminds the fact that the biological membranes contain various (protein) channels through which the molecules can pass (in a passive way, due to concentration difference, or in an active way, with a consumption of energy), in a rather selective manner. However, the fact that the communication of objects from a compartment to a neighboring compartment is controlled by the "reaction rules" is mathematically attractive, but not quite realistic from a biological point of view, that is why there were also considered variants where the two processes are separated: the evolution is controlled by rules as above, without target indications, and the communication is controlled by specific rules (by symport/antiport rules).

We have arrived in this way at the important feature of P systems, concerning the way of using the rules. The key phrase in this respect is: in the maximally parallel manner, non-deterministically choosing the rules and the objects.

More specifically, this means that we assign objects to rules, non-deterministically choosing the objects and the rules, until no further assignment is possible. More mathematically stated, we look to the set of rules, and try to find a multiset of rules, by assigning multiplicities to rules, with two properties: (i) the multiset of rules is applicable to the multiset of objects available in the respective region, that is, there are enough objects in order to apply the rules a number of times as indicated by their multiplicities, and (ii) the multiset is maximal, no further rule can be added to it (because of the lack of available objects).

Thus, an evolution step in a given region consists in finding a maximal applicable multiset of rules, removing from the region all objects specified in the left hand of the chosen rules (with the multiplicities as indicated by the rules and by the number of times each rule is used), producing the objects from the right hand sides of rules, and then

distributing these objects as indicated by the targets associated with them. If at least one of the rules introduces the dissolving action, then the membrane is dissolved, and its contents become part of the immediately upper membrane – provided that this membrane was not dissolved at the same time, a case where we stop in the first upper membrane which was not dissolved (at least the skin remains intact). [Păun 1998]

## Random Function

In common languages as C rand function is defined as a linear congruential generator. A linear congruential generator (LCG) represents one of the oldest and best-known pseudorandom number generator algorithms. The theory behind them is easy to understand, and they are easily implemented and fast.

The generator is defined by the recurrence relation:

$$X_{n+1} = (aX_n + c) \pmod{m}$$

where  $X_n$  is the sequence of pseudorandom values, and:

$0 < m$  the "modulus"

$0 < a < m$  the "multiplier"

$0 < c < m$  the "increment" (the special case of  $c = 0$  corresponds to Park Miller RNG)

$0 < X_0 < m$  the "seed" or "start value"

are integer constants that specify the generator.

While LCGs are capable of producing pseudorandom numbers, this is extremely sensitive to the choice of the coefficients  $c$ ,  $m$ , and  $a$ . [Bravo, 2002]

The most efficient LCGs have an  $m$  equal to a power of 2, most often  $m = 2^{32}$  or  $m = 2^{64}$ , because this allows the modulus operation to be computed by merely truncating all but the rightmost 32 or 64 bits. The following table lists the parameters of LCGs in common use, including built-in *rand()* functions in various compilers.

Source	m	a	c	output bits of seed in <i>rand()</i> / <i>Random(L)</i>
Numerical Recipes	$2^{32}$	1664525	1013904223	
Borland C/C++	$2^{32}$	22695477	1	bits 30..16 in <i>rand()</i> , 30..0 in <i>lrand()</i>
glibc (used by GCC)	$2^{32}$	1103515245	12345	bits 30..0
ANSI C: Watcom, Digital Mars, CodeWarrior, IBM VisualAge C/C++	$2^{32}$	1103515245	12345	bits 30..16
Borland Delphi, Virtual Pascal	$2^{32}$	134775813	1	bits 63..32 of ( <i>seed</i> * <i>L</i> )
Microsoft Visual/Quick C/C++	$2^{32}$	214013	2531011	bits 30..16
Apple CarbonLib	$2^{31} - 1$	16807	0	see Park-Miller RNG

Fig. 2. Relation between random function and LCG parameters for each compiler.

Historically, poor choices had led to ineffective implementations of LCGs. A particularly illustrative example of this is RANDU which was widely used in the early 1970s and resulted in many results that are currently in doubt because of the use of this poor LCG. Moreover, If a linear congruential generator is seeded with a character and then iterated once, the result is a simple classical cipher called an affine cipher; this cipher is easily broken by standard frequency analysis.

When using a LCG, the numbers are distributed into Hyperplanes. If a set of random numbers are part of the same Hyperplane, then the randomness is poor. This is what happens when we use LCGs to generate random numbers. See Fig 3.

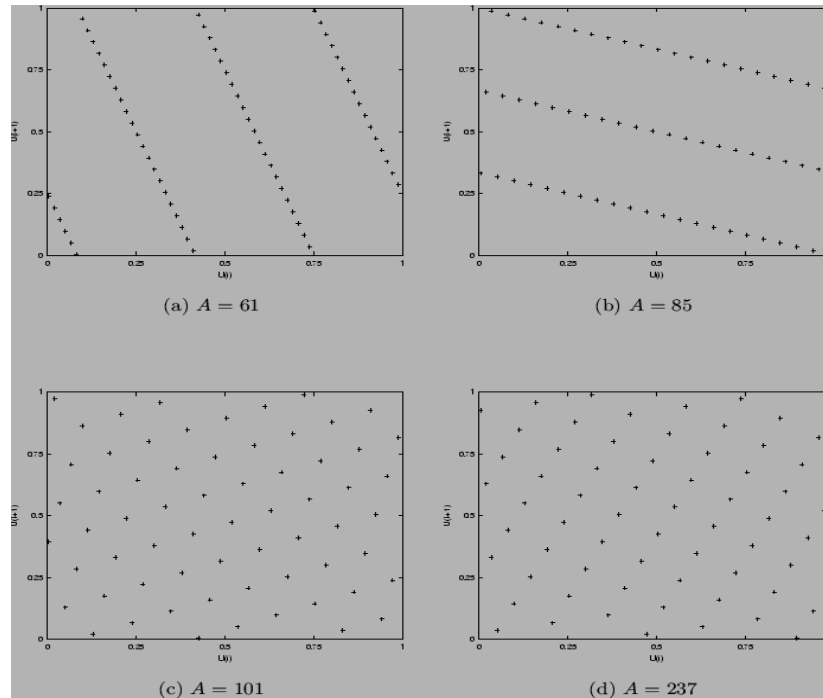


Fig. 3. Random distribution provided by LCG:  $X_{n+1} = (aX_n + c) \bmod m$

It seems that when using any of the compilers mentioned above, the determinism in our model practically does not exist. Moreover, when we fix a seed, we can totally reproduce the same sequence of numbers over and over again. In other words, the same seed generates the same output every time.

Poor randomness and predictability are signs of a deterministic process. This makes no reliable the idea of generating a non deterministic process as it is the one occurring within the living cells.

### Rules Applicability: Implementation of Non Determinism

Applying evolution rules in a p\_system is meant to be purely random. The way that reactions occur within the living cells is non deterministic. A common method to implement this behavior is to use the RND function. Nowadays, there are several methods of application of evolution rules which have been implemented. Algorithms as *Step by step Max applicability benchmark*, *Minimal applicability benchmark* [Fernandez,2006]. All of them study this point and try to improve the performance when applying rules. Here is an example of algorithm that applies the rules based on an applicability benchmark.

```

(1)  $\omega_{R(U)} \leftarrow \emptyset_{M(R(U))}$ 
(2) REPEAT
(3)  $r_i \leftarrow \text{random}(A)$ 
(4)  $\max \leftarrow \text{maximalApplicable}(r_i, \omega)$ 
(5) IF  $\max = 0$  THEN
(6)  $A \leftarrow A - \{r_i\}$ 
(7) ELSE BEGIN
(8)  $k \leftarrow \text{random}(1, \max)$ 
(9)  $\omega_{R(U)} \leftarrow \omega_{R(U)} + \{(r_i, k)\}$ 
(10)  $\omega \leftarrow \omega - k \cdot \text{input}(r_i)$ 
(11) END
(12) UNTIL  $|A| = 0$ 

```

Fig 4. Maximal applicability benchmark algorithm.

As shown, there are two calls to the random function. The current implementation of the random function makes the entire algorithm not very accurate on simulating the inherent non determinism within the living cells. The main reason is because the use of LCG which produces creates poor randomness and generates predictable output streams, while in a real scenario this should not occur.

**Rules Applicability: ICG, New Implementation Proposal**

In order to simulate randomness better, we must use more accurate random number generators.

The random generator we propose is able to simulate randomness in a better way. As the LCGs are proved not to be good for this simulation, We focus on the non linear ones.

The non linear congruential generator we propose here, is an Inversive congruential generators (ICGs).

Inversive congruential generators are a type of nonlinear congruential pseudorandom number generator, which use the modular multiplicative inverse [2] (if it exists) to generate the next number in a sequence. The standard formula for an inversive congruential generator is

$$X_{n+1} = \overline{(aX_n + c)} \pmod m$$

Sometimes the Parallel Hyperplanes phenomenon inherent in LCGs may cause adverse effects to certain simulation applications because the space between the hyperplanes will never be hit by any point of the generator, and the simulation result may be very sensitive to this kind of regularities. Inversive Congruential Generators (ICG) are designed to overcome this difficulty. It is a variant of LCG:

where  $\bar{c} = 0$  if  $c = 0$  and  $\bar{c} = c^{-1} \pmod M$ . To calculate  $\bar{c}$ , one can apply the reverse of Euclid's algorithm to find integer solutions for  $\bar{c}c + KM = 1$ .

Although the extra inversion step eliminates Parallel Hyperplanes (see Fig. 5), it also changes the intrinsic structures and correlation behaviors of LCGs. ICGs are promising candidates for parallelization, because unlike LCGs, ICGs do not have long-range autocorrelations problems.

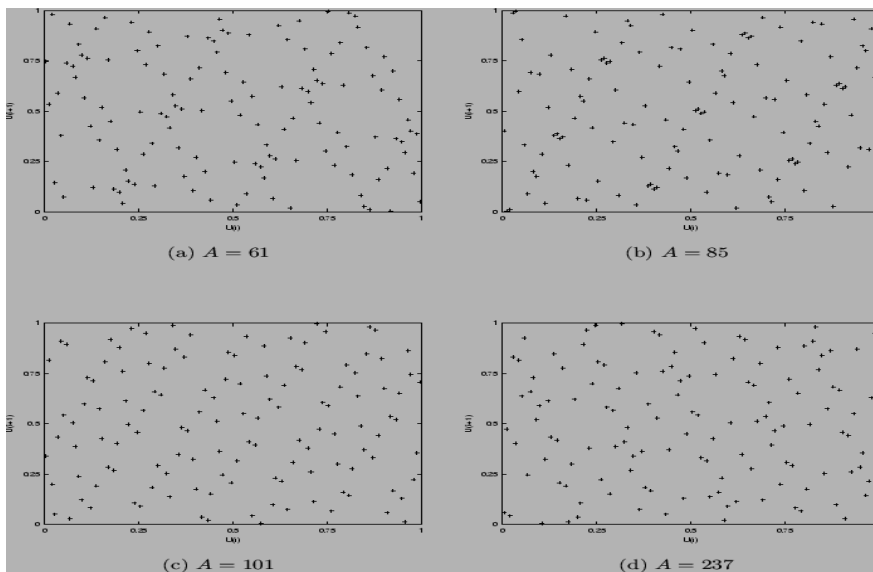


Fig. 5. Random distribution produced by ICG  $X_{n+1} = \overline{(aX_n + c)} \pmod m$

As shown the numbers are not distributed into Hyperplanes. This improves the simulation in terms of randomness which get us much closer to the idea of non determinism we are looking for.

Thus, the implementation of our random function to be used by our p-systems is:

Function  $RANDOM\_ICG(n) = (A * RANDOM\_ICG(n-1) + C) \bmod M$  where  $0 \leq n \leq M$

and M the number of evolution rules. The parameters we propose are:

A=237

M=  $2^{32}$

C=1265

$RANDOM(0)$  is the seed of the ICG and it can be set to any arbitrary number.

As shown in Figure 2 it is proved that this ICG does not generate parallel hyperplanes which get us closer to the idea of pure randomness in our model.

---

## Conclusion and Further Work

---

In this paper, we have studied some topics of membrane computing. As a part of this study, we have explained some concepts of the p-systems. Concepts such as:

1. Components
2. Interactions between the components.
3. The evolution of a p-system.

Moreover, we have focused our work on a specific part of the p-systems: Evolution rules application. The way that rules are applied in a region must be purely random. In order simulate this behavior we see that random function has been used by most developers. Most Compilers have implemented the random function by using LCG. After analyzing LCG we have concluded that it is a poor tool in terms of randomness and non determinism. As stated, is practically impossible to simulate a non deterministic process through a deterministic machine. However we can get closer to the idea of non determinism by increasing the quality of the random number generators.

By implementing and using a new random function we have been able to provide a better simulation in terms of randomness. This function uses the ISG we proposed in the above section. The random numbers generated by ICG are not placed in Parallel hyperplanes which improves simulation in terms of randomness.

Although it is practically impossible to simulate a non deterministic process by using deterministic tools as computers, we can improve the quality of simulation by using new random generators. This can be noticeable when the number of evolution rules increases within a given region. Although we approached the idea of randomness in the evolution rules application process, we still need to work on avoiding predictability as we could guess a given random number by knowing the initial value or seed of the ICG. [Blackburn, 2004] In the future, we will try to improve even more the simulation explained on this paper in terms of randomness and non determinism.

---

## Bibliography

---

- [Păun 1998] "Computing with Membranes", Journal of Computer and System Sciences, 61(2000), and Turku Center of Computer Science-TUCS Report n° 208, 1998.
- [Blackburn, 2004] "Predicting nonlinear pseudorandom number generators" Journal Mathematics of Computation. 74 (2005), 1471-1494.
- [Arroyo, 2001] "Structures and Bio-language to Simulate Transition P Systems on Digital Computers," Multiset Processing International Workshop Membrane Computing, Curtea de Arges (Romania), August 2002, Springer-Verlag, Vol 2597, pp. 19-32, Berlin, 2003
- [Bravo, 2002] " Una funcion random poco aleatoria". Spanish journal of physics , ISSN 0213-862X Vol.16 pp 60-62
- [Fernandez,2006] "New Algorithms for Application of Evolution Rules based on Applicability Benchmarks". BIOCAMP06: International Conference on Bioinformatics and Computational Biology, Las Vegas, (June, 2006)

---

## Authors' Information

---

*Alberto Arteta Albert* – Associate professor U.P.M Crtra Valencia km 7, Madrid-28031, Spain;  
e-mail: [aarteta@eui.upm.es](mailto:aarteta@eui.upm.es)

*Research: Membrane computing, Education on Applied Mathematics and Informatics*

*Luis Fernández Muñoz* – Associate professor U.P.M Crtra Valencia km 7, Madrid-28031, Spain;  
e-mail: [setillo@eui.upm.es](mailto:setillo@eui.upm.es)

*Fernando Arroyo Montoro*– Associate professor U.P.M Crtra Valencia km 7, Madrid-28031, Spain;  
e-mail: [farroyo@eui.upm.es](mailto:farroyo@eui.upm.es)

## MODELING LANGUAGE OF MULTI-AGENT SYSTEMS = PROGRAMMING TEMPLATE

Rubén Álvarez-González, Miguel Angel Díaz Martínez

*Abstract:* The modeling languages are designed to make easier the software development. That is why so many times they are included in the development methodology. In 2001 the OMG proposed model driver architecture for the software development (MDA). In this architecture are transformations which are used between the models to get others. The goal is to show a new way to develop applications using the Agents Oriented paradigm. To do it the MDA is showed and the methodology agent's models are studied. There is no methodology which uses the transformations between the models, so the meta-models group and the transformations between them should be researched.

*Keywords:* MDA, OMG, Modeling Languages, Meta-Models, Models, MAS, Agents, MDD, MDE.

*ACM Classification Keywords:* C.2.4 Distributed Systems - Distributed applications, D.2.11 Software Architectures – Languages.

---

### Introduction

---

The software development evolves all time. It starts when the programmers made the software using machine language. Now, the software with high level programmer languages is built. This evolution tries to make the development software with a language as similar as the human one instead to a machine language. All of these time different paradigms are proposed. The problems are analyzed in a way more natural with these paradigms.

Two of these paradigms are very similar. These are: Object Oriented (OO) and Agent Oriented (AO). The main entity in both paradigms, object to OO and agent to AO, encapsulate their state. The objects use private attributes to save their state and the agents save their state as beliefs. An object might interact with other objects using public method, the first one forces the second to do something when the first object calls for it on a public method. On the opposite, an agent sends messages with the other agent. The agents negotiate with these messages to do something, and an agent ever forces another agent to do something. In other words, these paradigms present a different view of the world. OO analyzes the world as an object group, these objects interact between them. AO interprets the world as an autonomous agent group which collaborates between them. [Bernon et al, 2005]

To develop the software using these paradigms work methodologies were proposed. These methodologies are used to guide the software development process. In the AO case some methodologies exist: Gaia, PASSI, ROADMAP, etc [Wooldridge et al, 2000], [Cossentino, 2005] and [Juan et al, 2002]. All of these use modeling languages to build new models. The solutions are represented with those models. A model is an affirmation collection, which is true or false, about a study system [Seidewitz, 2003].

This document's objective is to present a new way for software development using the AO paradigm. To do this the document has three main parts. The first section presents the Model Driven Architecture (MDA). In the second part the modeling languages which are used in the AO methodologies are studied. And in the third division the differences between the AO languages models and MDA technology are studied. The work finishes with the author conclusions.



---

## MDA

---

There are some problems in the software development. These appear when the developer group wants to integrate systems which exist already with new technologies [Kent, 2002]. The OMG want to resolve these problems and to do it this group proposed the model driver architecture in 2001.

MDA defines the IT systems (Information Technology) in two parts. The first one is the functionality specification, and the second part is the implement specification of the functionality in a particular technology. This definition is one of the main MDA characteristics. MDA uses platform independent models (PIM) and platform specific models (PSM). The most important advantages of these are [OMG, 2001]:

- To make easier to verify that a model is or is not correct.
- To make easier the generation of an implementation in a different platform with the same structure and behaviour.
- To make possible the definition, of integration mechanism and interoperability between systems in an independent way.

A PIM is a formal specification of the system structures and their functionality. This specification doesn't take into account the technical details. A PSM is a specification model of the system's platform. The relation between PIM and PSM is that defined functionality in the PIN will be executed on the platform which is specified in the PSM.

Like the source code, or like the natural language, the models and the transformations need a proper language for their representation. For the models, these languages are called "modeling languages" or "meta-models". At the same time, another kind of language is necessary to define the meta-models. These new languages are entitled "meta-modeling languages" or "meta-meta-models". The meta-model standard language is MOF [OMG, 2003].

The model driver architecture made up for meta-meta-models, meta-models and models. This architecture has three levels of meta-levels (Fig 1).

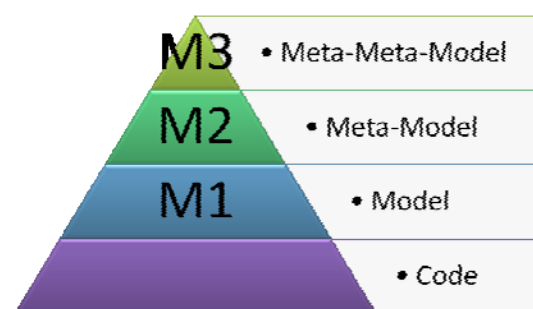


Figure 1. Model's Meta-levels

The IT professionals want to develop applications with models and some transformation between them. However, to build the applications in that way, it will be necessary to transform model to code. So it has three types of transformation: transformation between the same type of models (PIM to PIM), transformation between the different type of models (PIM to PSM or PSM to PIM) and transformation model to text (with this the model could be transformed into a code or into a model) [OMG, 2001].

---

## Agent Modeling Languages

---

The modeling languages are created to make software development easier. For this reason, they are incorporated in methodology development. The methodology of software development is a group made up of process, techniques and helps to make this type of asset.

Sometimes the methodology's models are described with a natural language. For example, MAS-CommonKADS methodology [Iglesias et al, 1998] uses this kind of description to describe the models and other aspects of it [Gómez Sanz, 2002]. In this way, the automatic analysis of a specification is more difficult [Gómez Sanz, 2002]. It is also possible to define the model with semi-formal languages like UML. UML's language makes feasible to automate the model's analysis, so a model could be checked.

UML is a modeling language or meta-model to develop software using on OO paradigm. In 2000, James Odell, H. Van Dyke Parunak and Bernahard Bauer proposed the AUML modeling language. AUML is an UML extension to MAS (Multi-Agent System) [Odell et al, 2000]. A modeling language's extension redefines the own language, taking off some elements and adding others. The goal of these changes is to adapt the modeling language to new needs.

This modeling language has three levels. Each layer consists of one or more models. In the first layer the protocols are specified as an interaction between roles. In the second layer the interactions between agents are defined. The last layer is used to specify the internal process of each agent.

A protocol is a tidy group of messages that are changed between two entities. In AUML the entities are the roles. An agent could have one or more roles.

To represent the interactions between agents, three different models are used: sequence diagram, collaboration diagram and activity diagram.

The sequence diagrams are used to represent a temporal sequence of messages between agents. With the collaboration diagram, it defines the sequence of messages between agents, but this sequence is not temporal. The activity diagram, which represents a sequence of messages, the sequence diagram are different because with the first one it is able to have an explicit control of the threads. This control is important to model complex models.

To specify the internal processes of agents, the sequence diagram and the state diagram are used. With the sequence diagram, the execution process orders are defined. The state diagram can also be used to specify an agent's process. To do it, the different states of an agent and their transitions are defined.

No Agent Oriented development methodologies use all AUML's models. There are methodologies which use some AUML models, an example is ROADMAP. Following that the models use in the principal AO methodologies are presented. The studied methodologies are: GAIA, RORADMAP, MESSAGE, INGEIAS, TROPOS and SODA.

Gaia can be the most influential methodology to analysis systems as an organization [Bernon et al, 2005]. The Gaia organization's made up of a group of roles which are assigned to an agent. In this methodology the models are used in analysis and in the design phase. These models are: interaction model and role model in the analysis phase and agent model, services model and relations model in the design phase. [Wooldridge et al, 2000]

In figure 2 you can appreciate the different relationships between models. In the interaction model the protocols are defined. A protocol is a relationship between roles, after they are specified in the role model. The agent model is used to identify the kind of system's agents. A kind of agent implements a role group each. The role's activities

are specified in the service model. A service might need one or more protocols. To finish, in the relations model, the agents are associated between them. It is possible to deduce the last model from the iterations model, roles model and agent model. The relations model's objective is to find blocks. These blocks can appear if some kind of agent has too big a task. [Wooldridge et al, 2000]

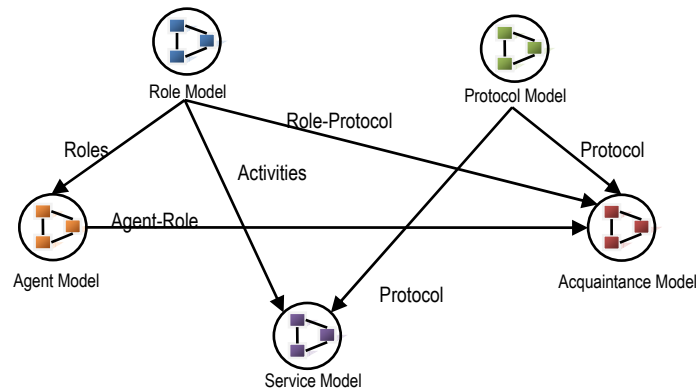


Figure 2. Relationship between GAIA's models [Wooldridge et al, 2000]

GAIA's methodology has a few disadvantages, which are resolved by ROADMAP [Juan et al, 2002]. This methodology modifies GAIA to improve the complex open system's process development.

The methodology ROADMAP has models to specify, analyze and design. Specification phase and analysis phase use a case model, environment model, knowledge model, roles model, protocol model and interaction model. In the design phase, it adds detail into the analysis phase's models until all necessary information for the system development is reflected. The previous phase also uses three new models: agent model, service model and acquaintance model. The last three models are extracted from GAIA.

The use-case model definition does not change in respect to its specification in UML. The author only modifies its interpretation. In UML, the user interacts with the software to work. In the new interpretation, the user interacts with a group of agents which have all the knowledge and functionalities required. This abstraction is used to make modeling of the problem easier.

The environment model is defined, with the use-case model's information, the different environment zones where the system will execute. These zones are hierarchically specified, at first a small group of zones are defined; later the sub-zones are specified and in this way until enough detail of the environment is presented. The relationship zones are modeled using UML's class diagram: hierarchical relationship, aggregation, etc.

From the two previous models the required knowledge can be deduced and the agent's behavior in each zone for each use-case. These knowledge and behaviors are represented in the acquaintance model.

In ROADMAP add hierarchy to the role model of GAIA. In this new model a father role can use all protocols for its sons and make their activities.

The ROADMAP's interaction model is not the GAIA's interaction diagram. This model is the AUML's interaction model, where the messages are exchanged between roles. The protocol model used in ROADMAP is the GAIA's protocol.

MESSAGE is a generic methodology for MAS development. This methodology is created because it was necessary for the telecommunication industry. To specify the system is need five view: organization view, goal and task view, agent/role view, interaction view and domain view.

To define all mentioned views in the previous paragraph the MESSAGES methodology uses eight nodes and four kind of relationship. Only one node is not a MAS's node, the name of this node is "resource". The resource node represents a non autonomy entity as data base or external programs. [AUML Web Site, 2003]

In the organization diagram the relationship between agents, organization, roles and resources are defined. These elements and their dependences are present in the goal/task view. The agent/role view is used to specify the goals and tasks to each pair of agent/role. Interaction model allows defining the interaction between roles. To finish, the domain view is used to define domain's necessary concepts to construct the system. Every model except one use a meta-model specifies for this methodology, only the domain view use a UML model, the class model.

In 2002 a doctoral thesis from Jorge J. Gómez Sand was introduced [Gómez Sanz, 2002]. This thesis proposed a new methodology which is called INGENIAS, and is for the MAS development. This methodology uses five meta-models for the system definition: organization meta-model, environment meta-model, agent meta-model, interaction meta-mode, goal/task meta-model. The different models have a relation between themselves using them entities, for example, an interaction model and an agent model have a connection if the agents from the agent model are in the interaction model.

The agents, the roles and the groups which make the system can be specified because of the organization's model. This model is based in the AALAADIN meta-model [Ferber et al, 1998]. In this one, an agent can be in one o more groups, and a group can have one or more roles. An agent which is inside a group only can have a role and a role only can have an assigned group. An organization is made up of groups and the agents from those groups.

The environment is defined thanks to the meta-models where the relationship between the agents, the groups, resources and applications are. The resources are in an agent or in a group.

In the goal/task model it is specified how the state of an agent changes in time. The task can change an agent's state. This modification can end the goal. An objective could be made up of the same sub-goals. In the same way, a task can be made up of some sub-jobs. So, an agent can execute a task's group which should help to end the group of objectives.

To define the relationship between the agents the interaction's model is used. This model can be represented for different meta-models: UML's diagram collaboration, GRASIAS's diagram interaction or the AUML's diagram protocol. The authors don't give their preference between any meta-models exposed.

To finish, the meta-model of the agents can specify the relationship between an agent in particular, the task which it can executed and the goals which can found. All of these meta-models have a connection doing a group which can make a whole system definition.

---

## Discussion

---

It can be shown that the methodologies which were presented in the last section are based, more or less, in the OMG technology. Some of these methodologies redefine the model's language. Other of these methodologies uses MOF to specify their own modeling languages. But none use the transformation that MDA proposes.

---

UML is a common visual language for the description of any component from an application which is going to be developed with an OO paradigm. In other words, UML can be considered like a language to define the implementation templates [Thomas, 2004]. So, if the meta-models from UML are used, the final goal does not change. These models only can add, delete components or modify the models interpretation.

On the other hand, if the models created are not complement with transformations which were exposed about MDA, the goal found for the language created will be the same as the goal discussed in the last paragraph.

The MDA's objective is the whole definition of an application which is independent of the technology or the paradigm chosen for the implementation. The PSM's models allow you to define the paradigms and technologies which must be used in each implementation. [OMG, 2001]

The modeling used in AO is a way to structure the computer application: data, algorithms, etc. All these elements can be implemented with PIM models. These models allow to the programmers to block out of hardware platform or software. With this, the design software is closer than before to human language. In other words, a developer can use as application's element a car; the developer does not need to know if this car is an object or an agent.

---

## Conclusion

In this work it is presented what model driven architecture is and its advantages. After each model used in the different methodology which was proposed for AO paradigm was studied.

AO paradigm is after OO paradigm. At first this one was programmed using OA paradigm. Later methodologies to make easier the software development with this paradigm appeared. Soon after, textual models were incorporated to the methodologies. At the end these add graphic models to themselves. The graphic models allow to have a visual communication and to have as well a better communication between requirement and implementation.

Some authors affirmed that AO is closer to owner world's view that OO is [Jennings, 2000]. Anyway AO is still based in variables and loops used (for, while, etc). *It* can also say that to develop applications with AO paradigm it is necessary to think in declare functions and variables instead use the natural environment's elements (bread, cars, etc.).

It is obvious that the methodologies and the models studied in this document have improved the AO applications development. But as it happen with the paradigms this is far from the way that a person understands the process which wants to automate.

To finish we can affirm that it is necessary to study the way to build a group of meta-models and the transformations to allow the model driven development uses AO paradigm.

---

## Bibliography

- [AUML Web Site, 2003] AUML Web Site Modeling Notation Source MESSAGE. In AUML Web Site. - AUML, Marzo 12, 2003. - Febrero 20, 2009. - <http://www.auml.org/auml/documents/>.
- [Bernon et al, 2005] C. Bernon, M. Cossentino and J. Pavón. An Overview of Current Trends in European AOSE Research. In Informatica. - Ljubljana : [s.n.], 2005. - Vol. 29. - pp. 379-390 .
- [Cossentino, 2005] M. Cossentino. From requirements to Code with the PASSI Methodology. In Agent-Oriented methodologies / ed. Henderson-Sellers B and Giorgini P. - [s.l.] : Idea Group Publishing, 2005.

- 
- [Ferber et al, 1998] F. Jacques and G. Olivier. A meta-model for the analysis and design of Organizations in multi-agent systems. In Third International Conference on Multi Agent Systems (ICMAS'98). - 1998. - pp. 128-135.
- [Gómez Sanz, 2002] J. Gómez Sanz. Multi-Agent System Modelling (In Spanish: MODELADO DE SISTEMAS MULTI-AGENTE). In Tesis Doctoral / Sistemas Informáticos y Programación. - Madrid : [s.n.], 2002.
- [Iglesias et al, 1998] C. Iglesias, M. Garijo, J. Gonzales, J. Velasco. Analysis and Design of Multiagent Systems using MAS-CommonKADS. In 4th International Workshop on Agent Theories, Architectures, and Languages.. - Londres : illustrated, 1998. - pp. 313-328.
- [Jennings, 2000] N. Jennings. On agent-based software engineering. In Artificial Intelligence. - [s.l.] : Elsevier Science B.V., 2000. - 117. - pp. 277–296.
- [Juan et al, 2002] T. Juan, A. Pearce and L. Sterling. ROADMAP: Extending the Gaia Methodology for Complex Open Systems [Conference] // First International Joint Conference on Autonomous Agents & Multi-Agent systems. - [s.l.] : ACM Press, 2002. - pp. 3-10.
- [Kent, 2002] S. Kent. Model Driven Engineering. In Proceedings of the Third International Conference on Integrated Formal Methods / ed. Butler M, Petre L and Sere K. - Turku : [s.n.], 2002. - pp. 286-298.
- [Odell et al, 2000] J. Odell, H. Van dyke Parunak and B. Bauer. Extending UML for Agents. In Proc. of the Agent-Oriented Information Systems, Workshop at the 17th National conference on Artificial Intelligence. - Austin : [s.n.], 2000.
- [OMG, 2003] OMG The Object Management Group (OMG). OMG's MetaObject Facilit. – OMG. In - <http://www.omg.org/docs/formal/02-04-03.pdf> . Abril 02, 2003. - 1.4. - Enero 07, 2009.
- [OMG, 2001] OMG The Object Management Group (OMG). MDA Specification. - OMG, 2001. - 1.0.1.
- [Seidewitz, 2003] E. Seidewitz. What the model's mean?. In IEEE Software. - 2003. - Vol. 20. - pp. 26-32.
- [Thomas, 2004] D. Thomas. MDA: Revenge of the Modelers or UML Utopia? In IEEE SOFTWARE. - [s.l.] : I E E E Computer Society, 2004. - p. 3.
- [Wooldridge et al, 2000] M. Wooldridge, N. Jennings and D. Kinny. The Gaia Methodology for Agent-Oriented Analysis and Design. In Journal of Autonomous Agents and Multi-Agent Systems. – 2000, Vol. 3, Num.3 - pp. 285-312.
- [Shannon, 1949] C. Shannon. The Mathematical theory of communication. In: The Mathematical Theory of Communication. Ed. C.E.Shannon and W.Weaver. University of Illinois Press, Urbana, 1949.
- 

### Authors' Information

---

*Rubén Álvarez-González* – Student of Natural Computing Group. Faculty of Computer Science. Technique University of Madrid. [ruben.alvarez.gonzalez@gmail.com](mailto:ruben.alvarez.gonzalez@gmail.com)

*Miguel Angel Díaz Martínez* – Professor of Computer Science School. Technique University of Madrid. [mdiaz@eui.upm.es](mailto:mdiaz@eui.upm.es)

---

## COMPARISON OF DISCRETIZATION METHODS FOR PREPROCESSING DATA FOR PYRAMIDAL GROWING NETWORK CLASSIFICATION METHOD

Ilia Mitov, Krassimira Ivanova, Krassimir Markov,  
Vitalii Velychko, Peter Stanchev, Koen Vanhoof

*Abstract:* This paper presents a comparison of four representative discretization methods from different classes to be used with so called PGN-classifier which deals with categorical data. We examine which of them supplies more convenient discretization for PGN Classification Method. The experiments are provided on the base of UCI repository data sets. The comparison tests were provided using an experimental classification machine learning system "PaGaNe", which realizes Pyramidal Growing Network (PGN) Classification Algorithm. It is found that in general, PGN-classifier trained on data preprocessed by Chi-merge achieve lower classification error than those trained on data preprocessed by the other discretization methods. The comparison of PGN-classifier, trained with Chi-merge-discretizator with other classifiers (realized in WEKA system) shows good results in favor of PGN-classifier.

*Keywords:* Data Mining, Machine Learning, Discretization, Data Analysis, Pyramidal Growing Networks

---

### 1. Introduction

---

Building of self-structured systems had been proposed to be realized on basis of special kind of neural networks with hierarchical structures, named as "growing pyramidal networks" (GPN) [Gladun, 2008]. Pyramidal network is a network memory, automatically tuned into the structure of incoming information. Unlike the neuron networks, the adaptation effect is attained without introduction of a priori network excess. The research done on complex data of great scope showed high effectiveness of application of growing pyramidal networks for solving analytical problems. Such qualities as simplicity of change incoming data, combining processes of information input with processes of classification and generalization, high associability makes growing pyramid networks an important component of forecasting and diagnosing systems [Gladun, 2003].

A realization of the growing pyramidal networks by the multidimensional numbered information spaces for memory structuring in the self-structured systems was presented in [Mitov et al, 2009]. The main advantage of the numbered information spaces is the possibility to build growing space hierarchies of information and the great power for building interconnections between information elements stored in the information base. Practically unlimited number of dimensions and the opportunity of representing and storing the information only about the existing parts of the knowledge make possible creating effective and useful tools [Markov, 2004].

To make difference, the new network model was named Pyramidal Growing Network (PGN). A classification machine learning system "PaGaNe", which realizes Pyramidal Growing Network (PGN) Classification Algorithm, based on the multidimensional numbered information spaces for memory structuring is realized. PGN Classification algorithm combines generalization possibilities of Propositional Rule Sets with answer accuracy like K-Nearest Neighbors. PGN is aimed to process categorical data. To extend possibilities of PaGaNe system in direction to work with nominal data a specialized tools for discretization are realized.

Discretization process is known to be one of the most important data preprocessing tasks in data mining.

Many machine learning techniques can be applied only to data sets composed of categorical attributes but a lot of data sets include continuous variables. One solution to this problem is to partition numeric variables into a

---

number of sub-ranges and treat each such sub-range as a category. This process of partitioning continuous variables into categories is usually termed discretization. The advantages of data discretization can be founded in different directions:

- the experts usually describe parameters using linguistic terms instead of an exact value. In a sense the discretization provides better perceiving of attributes;
- it provides regularization because it is less prone to variance in estimation from small fragmented data;
- the amount of data can be greatly reduced because some redundant data can be identified and removed;
- it provides better performance for the rule extraction.

Primary methods are:

- *Supervised or Unsupervised* [Dougherty et al, 1995]: In the unsupervised methods, continuous ranges are divided into sub-ranges by the user specified parameter – for instance, equal width (specifying range of values), equal frequency (number of instances in each interval), clustering algorithms like k-means (specifying number of clusters). These methods may not give good results in cases where the distribution of the continuous values is not uniform, where outliers affect the ranges significantly. Of course if no class information is available, unsupervised discretization is the sole choice. In supervised discretization methods class information is used to find the proper intervals caused by cut-points. Different methods have been devised to use this class information for finding meaningful intervals in continuous attributes. Supervised discretization can be further characterized as *error-based*, *entropy-based* or *statistics-based* according to whether intervals are selected using metrics based on error on the training data, entropy of the intervals, or some statistical measure.
- *Hierarchical or Non-hierarchical*: Hierarchical discretization selects cut points in an incremental process, forming an implicit hierarchy over the value range. The procedure can be *split* or (and) *merge* [Kerber 1992]. Some methods are non-hierarchical: for instance these, which scan the ordered values only once, sequentially forming the intervals.
- *Top-down or Bottom-up* or in other means *Split or Merge* [Hussain et al, 1999]: Top-down methods start with one interval and split intervals in the process of discretization. Bottom-up methods start with the complete list of all the continuous values of the feature as cut-points and remove some of them by "merging" intervals as the discretization progresses. Different thresholds for stopping criteria are used.
- *Static or Dynamic*: The static approach discretization is done prior to the classification task (in pre-processing phase). A dynamic method would discretize continuous values when a classifier is being built, such as in C4.5 [Quinlan, 1993]. Dynamic methods are mutually connected with corresponded classification method, which algorithm can work with real attributes.
- *Parametric or Non-parametric*: Parametric discretization requires input from the user, such as the maximum number of discretized intervals. Non-parametric discretization only uses information from data and does not need input from the user.
- *Global or Local* [Dougherty et al, 1995]: A local method would discretize in a localized region of the instance space (i.e. a subset of instances) while a global discretization method uses the entire instance space to discretize. So, a local method is usually associated with a dynamic discretization method in which only a region of instance space is used for discretization.
- *Univariate or Multivariate* [Bay, 2000]: Univariate discretization quantifies one continuous feature at a time while multivariate discretization considers simultaneously multiple features.



In our experiments we are focused on some representatives of unsupervised and supervised methods. From supervised methods we have chosen two methods, which are different from point of view of hierarchical directions and of forming intervals criteria. The first is Fayyad-Irani top-down method, which is based on the optimizing of local measure of entropy and as stopping criterion the Minimum Description Length (MDL) principle is used [Fayyad, Irani, 1993]. The second is Chi-merge – a bottom-up method based on chi-square statistics measure.

In section two, discretization methods, which we choose for realization in the experimental system PaGaNe, are described. Section three contains description of the program realization. Section four is aimed to represent some experimental results of classification, based on several benchmark training sets and comparison of accuracy of PGN-classifier trained with different discretization methods. In this section also is presented the comparison of the accuracy level of PGN-classifier, trained on the Chi-merge with other classifiers. As comparative space Weka system – Waikato Environment for Knowledge Analysis (Weka) [Witten, Frank, 2005] is used. Finally, conclusions and future work are presented.

---

## 2. Discretization Methods

---

We have chosen discretization methods from different classes in order to examine which of them supplies more convenient discretization for PGN Classification Method.

- *Equal Width Discretization* – the simplest unsupervised discretization method, which determines the minimum and maximum values of the discretized attribute and then divides the range into the user-defined number of equal width discrete intervals. There is no "best" number of bins, and different bin sizes can reveal different features of the data. Some theoreticians have attempted to determine an optimal number of bins.

- *Equal Frequency Discretization* – the unsupervised method, which divides the sorted values into  $k$  intervals so that each interval contains approximately the same number of training instances. Thus each interval contains  $n/k$  (possibly duplicated) adjacent values.  $k$  is a user predefined parameter.

- *Fayyad-Irani Discretization method* [Fayyad, Irani, 1993] – supervised hierarchical split method, which use the class information entropy of candidate partitions to select boundaries for discretization. Class information entropy is a measure of purity and it measures the amount of information which would be needed to specify to which class an instance belongs. It considers one big interval containing all known values of a feature and then recursively partitions this interval into smaller subintervals until MDL criterion or an optimal number of intervals is achieved.

The MDL Principle states that the best hypothesis is the one with minimal description length. As partitioning always decreases the value of the entropy function, considering the description lengths of the hypotheses allows balancing the information gain and eventually accepting the null hypothesis. Performing recursive bipartitions with this criterion leads to a discretization of the continuous explanatory attribute at hand. Fayyad-Irani Discretizator evaluates as a candidate cut point the midpoint between each successive pair of the sorted values. For each evaluation of a candidate cut point, the data are discretized into two intervals and the resulting class information entropy is calculated. A binary discretization is determined by selecting the cut point for which the entropy is minimal amongst all candidate cut points. This binary discretization is applied recursively, always selecting the best cut point. A MDL criterion is applied to decide when to stop discretization. It has been shown that optimal cut points for entropy minimization must lie between examples of different classes.

- *Chi-merge* [Kerber, 1992] – supervised hierarchical bottom-up (merge) method that locally exploits the chi-square criterion to decide whether two adjacent intervals are similar enough to be merged;

Chi-square ( $\chi^2$ ) is a statistical measure that conducts a significance test on the relationship between the values of a feature and the class. Kerber argues that in an accurate discretization, the relative class frequencies should be fairly consistent within an interval but two adjacent intervals should not have similar relative class frequency. The  $\chi^2$  statistic determines the similarity of adjacent intervals based on some significance level. It tests the hypothesis that two adjacent intervals of a feature are independent of the class. If they are independent, they should be merged; otherwise they should remain separate.

The bottom-up method based on chi-square is ChiMerge. It searches for the best merge of adjacent intervals by minimizing the chi-square criterion applied locally to two adjacent intervals: they are merged if they are statistically similar. The stopping rule is based on a user-defined Chi-square threshold to reject the merge if the two adjacent intervals are insufficiently similar. No definite rule is given to choose this threshold.

---

### 3. Software Realization

---

We have realized the described in section two discretization methods in the experimental system PaGaNe, which presents Pyramidal Growing Network (PGN) Classification Method, based on the multidimensional numbered information spaces for memory structuring [Mitov et al, 2009].

The user can choose one of described methods. For some of them additional parameters have to be pointed:

- *Equal Width*: The system gives the possibility the number of intervals  $k$  for the set of  $n$  instances, where  $r_{\min}$  and  $r_{\max}$  are respectively minimal and maximal values of the instances, to be:
  1. given by the user;
  2. calculated on the base of Sturges' Formula [Sturges, 1926]:  $k = \lceil \log_2 n + 1 \rceil$  (" $\lceil \cdot \rceil$ " denotes ceiling function). Using of this formula directly was observed not good results in the preliminary experiments because of big partitioning of the space, so we reduce twice the suggested result;
  3. calculated by Scott's Formula [Scott, 1979]:  $k = \left\lceil \frac{r_{\max} - r_{\min}}{h} \right\rceil$ ,  $h = \frac{3.5 * \sigma}{\sqrt[3]{n}}$ , where  $\sigma$  is the standard deviation;
  4. calculated on the base of Freedman-Diaconis rule [Freedman, Diaconis, 1981]:  $k = \left\lceil \frac{r_{\max} - r_{\min}}{h} \right\rceil$ ,  $h = \frac{2 * IQR}{\sqrt[3]{n}}$ , where  $IQR$  is interquartile range in the set.

When the user gives the number of intervals, as well as when this number is chosen using the Sturges-based formula, this number is applied for all real attributes. Scott's and Freedman-Diakonis formulas take account of distribution of each attribute and give different number of intervals.

- *Equal Frequency*: Here the user has to choose the number of intervals.
- *Fayyad-Irani*: This method use as stopping criteria MDL Principle, which does not need additional parameters.
- *Chi-merge*: The stopping rule is based on a Chi-square threshold, which depends of degrees of freedom (in our case – the number of possible values of class minus one) and the significance level (commonly used significance levels are 90%, 95%, 99%). The chi-square threshold table in the system is given from [Bramer, 2007].

In the pre-processing phase, as a result of implementing of chosen discretization method, the system builds a mapping function for the real values of each attribute to a number that correspond to the interval in which the value belongs to.

Figure 1 is a screenshot of the screen of the experimental system "PaGaNe", which visualize the results of discretization process using "Chi-merge" with parameter 90% significance level for attribute "sepal length" for "Iris" dataset from UCI repository [Asuncion, Newman, 2007]. Forming of five intervals and distribution of different class values in the intervals are seen.

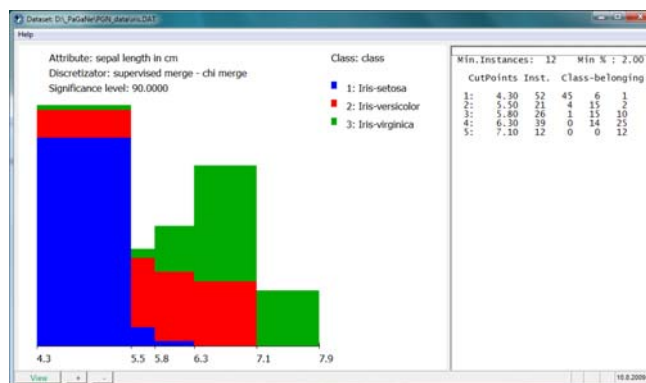


Figure 1. A Screenshot of visualizing discretization of attribute "sepal length in cm" of Iris database using Chi-merge discretizator in experimental system "PaGaNe".

In the right of the screen is shown the cut-points from each interval, number of instances of learning set and corresponded belonging to the class values of these instances.

The system uses these intervals to find the corresponded nominal values for real attributes in learning and examining sets. This converting of real data to categorical values gives the opportunity of PGN-classifier to be implemented on databases with the real values of attributes.

#### 4. Experimental Results

We have provided series of experiments with different datasets from UCI Machine Learning Repository [Asuncion, Newman, 2007]. The datasets Ecoli, Glass, Indian Diabetes, Iris, and Wine contain only real attributes. The datasets Forestfires, Hepatitis, Statlog contain real and categorical attributes. The original dataset Forestfires contains real numbers as class values (the burned area in the forest in ha) which is inconvenient for many classifiers. Because of this we replace positive numbers with "Yes" and zero numbers with "Not" depending of existing of fire or not.

The proportions of splitting the datasets to learning and examining sub-sets were respectively 2:1 (66.67%) and 3:1 (75%).

The realized discretizators were tested using different parameters: Chi-merge was examined with 90%, 95% and 99% significance level; Equal Width was controlled with supposed formulas for automatic defining of the number of intervals (Sturges, Scott, Freedman-Diaconis). The number of intervals for Equal Frequency Discretizator we gave the same as defined in Sturges formula. Fayyad-Irani is a non-parametric method.

In the table 1 and figure 2 the results are outlined.

The analysis of the received results shows that Chi-merge discretization method gives stable good recognition accuracy for PGN-classifier. Fayyad-Irani method gives in some cases very good results, but fails in other databases. Equal Frequency Discretizator gives relatively steady but not very good results. Instead of the fact that Equal Width Discretizator is the simplest one, it shows relatively good results and can also be used for discretization as pre-processor for PGN-classifier.

The main conclusion is: Chi-merge discretization method is more efficient for PGN-classifier than other methods.

Table 1. Comparison of accuracy answers of PGN-classifier trained with different discretization methods using several datasets.

Database	Ecoli	Ecoli	Forest fires	Forest fires	Glass	Glass	Hepatitis	Hepatitis	Indian Diabetes	Indian Diabetes	Iris	Iris	Statlog	Statlog	Wine	Wine
	2:1	3:1	2:1	3:1	2:1	3:1	2:1	3:1	2:1	3:1	2:1	3:1	2:1	3:1	2:1	3:1
Chi-merge:90%	82.14	77.38	61.63	51.94	77.46	64.15	82.35	76.32	73.44	68.75	96.00	94.59	84.78	84.30	96.61	100.00
Chi-merge:95%	83.04	79.76	56.98	55.04	74.65	71.70	82.35	71.05	74.61	72.92	94.00	94.59	85.65	83.72	91.53	97.73
Chi-merge:99%	78.57	80.95	58.14	51.94	74.65	73.58	86.27	76.32	75.00	71.88	96.00	91.89	84.78	84.30	93.22	100.00
Fayyad-Irani	70.54	28.57	57.56	54.26	38.03	56.60	84.31	78.95	74.61	72.92	92.00	94.59	85.22	84.88	96.61	95.45
Equal Frequency	74.11	73.81	56.40	51.94	70.42	69.81	84.31	76.32	75.39	66.67	88.00	94.59	84.78	83.72	91.53	97.73
Equal Width:Fr.-Diac.	74.11	78.57	51.16	57.36	61.97	58.49	84.31	78.95	74.22	69.79	70.00	91.89	85.65	84.30	93.22	95.45
Equal Width:Scott	79.46	75.00	56.40	58.91	59.15	75.47	84.31	76.32	78.52	69.27	96.00	97.30	84.78	83.14	93.22	97.73
Equal Width:Sturges	74.11	78.57	59.88	51.16	61.97	77.36	86.27	73.68	76.95	70.31	90.00	94.59	86.52	82.56	93.22	95.45

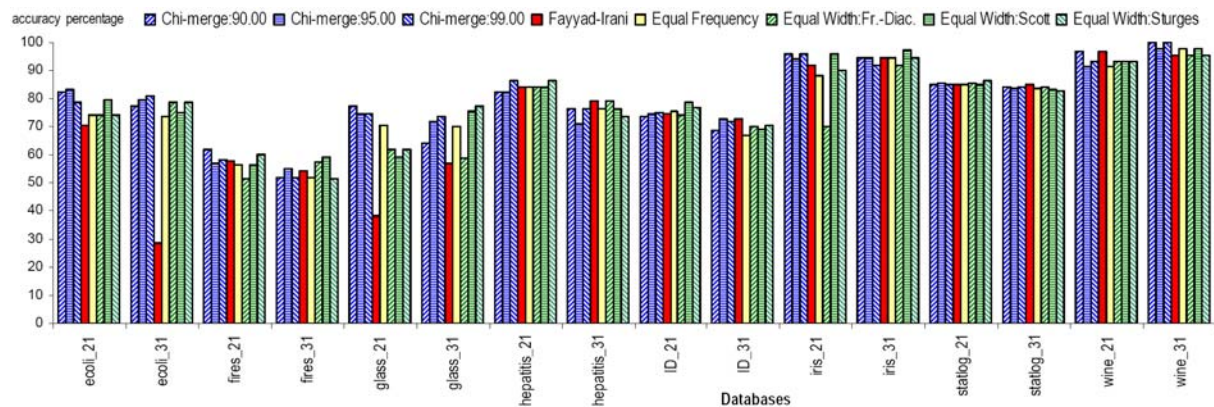


Figure 2. Graphical representation of the percentage of correct answers of PGN-classifier trained on data preprocessed by different discretization methods.

We compare the accuracy of PGN-classifier, trained with Chi-merge pre-processing discretization method (90% significance level) with other classifiers, realized in Waikato Environment for Knowledge Analysis (Weka) [Witten, Frank, 2005]. The software of Weka system can be obtained from <http://www.cs.waikato.ac.nz/ml/weka/>. We compare results achieved by PaGaNe with the results of the experiments with some algorithms in Weka, using the same datasets. We used classifiers, representatives of different recognition models:

- JRip – implementation a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER);
- OneR – one-level decision tree expressed in the form of a set of rules that all test one particular attribute;
- J48 – a Weka implementation of C4.5 that produces decision tree;
- IBk – k-nearest neighbor classifier;
- KStar – an instance-based classifier that uses an entropy-based distance function.

Table 2. Comparison of accuracy answers of PGN-classifier trained with Chi-merge discretizer with other classification methods, tested for databases, which contains numerical attributes.

Database	Learning Set : Examining Set split proportion	PGN+Chi	JRip	OneR	J48	IBk	KStar
Ecoli	66.67%	82.14	80.36	60.71	76.78	79.46	79.46
Ecoli	75%	77.38	89.29	61.90	77.38	80.95	80.95
Forestfires	66.67%	61.63	61.05	56.40	60.46	62.79	57.56
Forestfires	75%	51.94	44.18	53.49	51.16	51.94	58.14
Glass	66.67%	77.46	53.52	57.75	64.79	70.42	73.24
Glass	75%	64.15	68.71	57.81	70.31	71.88	73.44
Hepatitis	66.67%	82.35	82.69	86.54	84.61	75.00	78.85
Hepatitis	75%	76.32	76.92	71.79	71.79	69.23	66.67
Indian_Diabetes	66.67%	73.44	75.00	71.09	71.48	73.83	71.09
Indian_Diabetes	75%	68.75	75.00	71.88	79.17	71.35	70.83
Iris	66.67%	96.00	98.00	96.00	98.00	98.00	98.00
Iris	75%	94.59	97.30	91.89	91.89	97.30	97.30
Statlog	66.67%	84.78	83.91	83.91	84.78	80.43	80.00
Statlog	75%	84.30	84.30	84.30	82.56	77.91	80.81
Wine	66.67%	96.61	84.75	86.44	88.14	96.61	96.61
Wine	75%	100.00	86.36	68.18	90.91	88.64	93.18

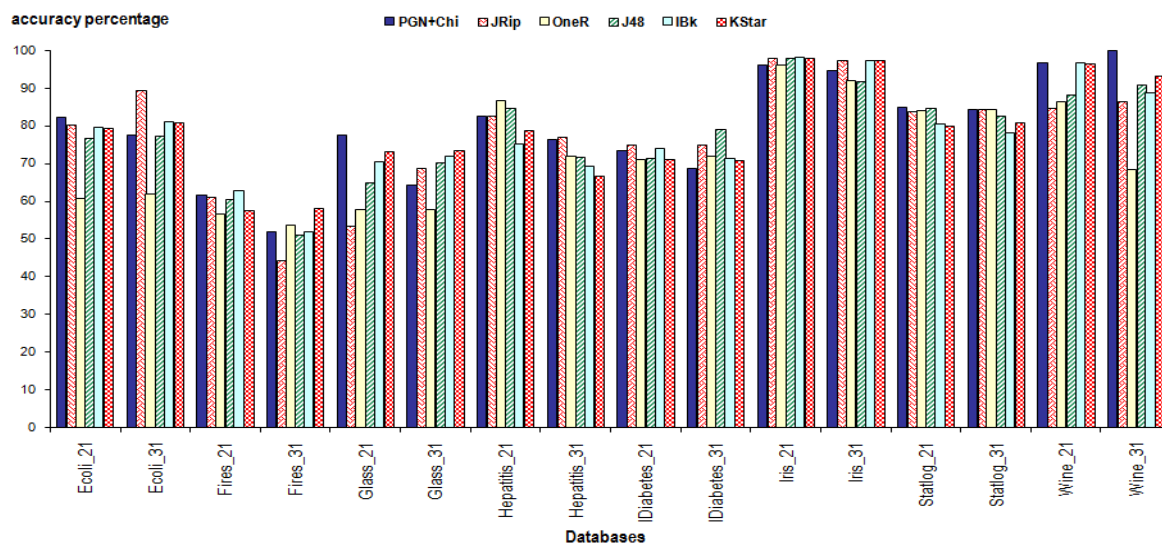


Figure 3. Comparison of PGN-classifier, pre-processed with Chi-merge discretization method with other classification methods, tested for databases, which contains numerical attributes.

Figure 3 illustrates the comparison of PGN-classifier, pre-processed with Chi-merge discretization method with WEKA classification methods, tested for databases, which contain numerical attributes. PGN-classifier in six cases is the best and in the all other cases is at the leading position.

---

## 5. Conclusion

---

A comparison of four representative discretization methods from different classes to be used with so called PGN-classifier which deals with categorical data was outlined in this paper. The main goal was to examine which of them supplies more convenient discretization for PGN Classification Method.

It was found that in general PGN-classifier trained on data preprocessed by Chi-merge achieves lower classification error than those trained on data preprocessed by the other discretization methods. The main reason for this is that using Chi-square statistical measure as criterion for class dependency in adjacent intervals of a feature leads to forming good separating which is convenient for the PGN-classifier.

The comparison of PGN-classifier, trained with Chi-merge-discretizator with other classifiers has shown good results in favor of PGN-classifier.

The achieved results are good basis for further work in this area. It is oriented toward realization of a new discretization algorithm and program tools, which will integrate the possibilities of already realized methods with specific features of PGN Classification Algorithm.

---

## Acknowledgements

---

This work is partially financed by Bulgarian National Science Fund under the project D 002-308 / 19.12.2008 "Automated Metadata Generating for e-Documents Specifications and Standards" and under the joint Bulgarian-Ukrainian project D 002-331 / 19.12.2008 "Developing of Distributed Virtual Laboratories Based on Advanced Access Methods for Smart Sensor System Design".

---

## Bibliography

---

- [Asuncion, Newman, 2007] A. Asuncion, D.J. Newman. UCI Machine Learning Repository. University of California, Irvine, CA, School of Information and Computer Science, <http://archive.ics.uci.edu/ml/> visited on 01.08.2009
- [Bay, 2000] S. Bay. Multivariate discretization of continuous variables for set mining. In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2000), pp. 315–319.
- [Bramer, 2007] M. Bramer. Principles of Data Mining. Springer Verlag London Limited 2007, ISBN-13: 978-1-84628-765-7.
- [Dougherty et al, 1995] J. Dougherty, R. Kohavi, M. Sahami. Supervised and unsupervised discretization of continuous features. In Proceedings of the 12th International Conference on Machine Learning (1995), pp. 194-202
- [Fayyad, Irani, 1993] U.Fayyad, K.Irani. Multi-interval Discretization of Continuous-Valued Attributes for Classification Learning. Proceedings of the 13th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, San Mateo, CA, 1993, pp.1022-1027.
- [Freedman, Diaconis, 1981] D. Freedman, P. Diaconis. On the Histogram as a Density estimator:  $L_2$  Theory. Probability Theory and Related Fields (Heidelberg: Springer Berlin) 57 (4): (December 1981), pp. 453–476
- [Gladun, 2003] V. P. Gladun. Intelligent Systems Memory Structuring. Int. Journal "Information Theories and Applications", Vol.10, No.1, 2003, pp. 10-14.
- [Gladun, 2008] V. Gladun, V.Velychko, Y. Ivaskiv. Selfstructured Systems. International Journal "Information Theories and Applications ", Vol.15, Number 1, 2008 pp. 5-13.
- [Hussain et al, 1999] F. Hussain, H. Liu, Ch. L. Tan, M. Dash. Discretization: An Enabling Technique. Technical Report – School of Computing, Singapore, June 1999.



- 
- [Kerber, 1992] R. Kerber. Discretization of Numeric Attributes. Proceedings of the Tenth National Conference on Artificial Intelligence, MIT Press, Cambridge, MA, 1992, pp.123-128.
- [Markov, 2004] K. Markov. Multi-Domain Information Model. Int. Journal "Information Theories and Applications", Vol.11, No.4, 2004, pp. 303-308.
- [Mitov et al, 2009] I. Mitov, Kr. Ivanova, Kr. Markov, V. Velychko, K. Vanhoof, and P. Stanchev. "PaGaNe" – A Classification Machine Learning System Based on the Multidimensional Numbered Information Spaces. "Intelligent Systems and Knowledge Engineering", 27-28.11.2009, Hasselt, Belgium (in appear).
- [Quinlan, 1993] J.Quinlan.C4.5: Programs for Machine Learning. M. Kaufmann, San Mateo, CA, 1993.
- [Scott, 1979] D. Scott. On Optimal and Data-based Histograms. Biometrika 66 (3), 1979, pp. 605–610.
- [Sturges, 1926] H. Sturges. The Choice of a Class Interval. J. American Statistical Association: 1926, pp. 65–66.
- [Witten, Frank, 2005] I. Witten, E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
- 

### Authors' Information

---

*Iliia Mitov* – PhD Student of the Institute of Mathematics and Informatics, BAS, Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: [mitov@foibg.com](mailto:mitov@foibg.com)

*Krassimira Ivanova* – Researcher; Institute of Mathematics and Informatics, BAS, Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: [kivanova@math.bas.bg](mailto:kivanova@math.bas.bg)

*Krassimir Markov* – Assoc. Professor; Institute of Mathematics and Informatics, BAS, Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: [markov@foibg.com](mailto:markov@foibg.com)

*Vitalii Velychko* – Doctoral Candidate; V.M.Glushkov Institute of Cybernetics of NAS of Ukraine, Prosp. Acad. Glushkov, 40, Kiev-03680, Ukraine; e-mail: [glad@aduis.kiev.ua](mailto:glad@aduis.kiev.ua)

*Peter Stanchev* – Professor, Kettering University, Flint, MI, 48504, USA / Institute of Mathematics and Informatics – BAS; Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: [pstanche@kettering.edu](mailto:pstanche@kettering.edu)

*Koen Vanhoof* – Professor, Universiteit Hasselt; Campus Diepenbeek; Department of Applied Economic Sciences; Research Group Data Analysis & Modelling. Wetenschapspark 5; bus 6; BE-3590 Diepenbeek; Belgium; e-mail: [koen.vanhoof@uhasselt.be](mailto:koen.vanhoof@uhasselt.be)

## MULTILINGUAL OPERA SUBTITLING EXCHANGE BETWEEN PRODUCTION AND BROADCASTER COMPANIES

Jesús Martínez Barbero, Manuel Bollain Pérez

*Abstract: Subtitling is used extensively by broadcasters both for foreign-language subtitling and as an access service to help people with a disability access television programs. The European Broadcaster Union (EBU) has create a working group in 2009-01-21 for use the standard Distribution Format Exchange Profile [1] (DFXP) for exchange subtitling information in XML with the Material Exchange Format [2] (MXF) file format. At this moment, in order to help the deaf people, broadcaster produces textual data in a file with a time-code referred of the picture. Subtitle live events require speech recognition or special keyboards where the words are presented as a union of several keys pressed. In this paper, we present an initiative based in B2B standards for the exchange, production, and broadcast multilingual subtitle for live Opera production.*

*Keywords: MPEG2, Multimedia, DTV, Subtitling, Broadcasting.*

---

### 1. Introduction

---

The organization and structure of TV broadcasting has three major divisions: production, distribution, and exhibition or diffusion. Production is usually performed by mobile units in the same location where the event takes place. The signal is usually uplinked to a satellite so that the downlink can be performed by any of the channels owning the broadcasting rights for each territory. This is a one-to-many transmission. The broadcasters downlink the satellite signal and they package it with the playout system of their channels to broadcast it through the Media owning the rights. Generally this transmission is one-to-many. Typically, the rights of a live TV event are sold to a broadcaster company for each country.

Opera event has inconveniences in the traditional TV broadcasting; most of the people don't understand operas' concerts and difficulty of understanding the words, international transmission adds another inconvenience: language.

With this work, we propose a system for subtitling opera's live events including translating processes without the necessity of live subtitling in each broadcaster headquarters.

This paper is organized as follows. Section 2 surveys the different standards and state of art for subtitling and data transmission used in the section 3 where model is presented. Section 4 concludes the paper by describing the benefits.

---

### 2. Overview

---

We start this section by providing a brief overview of the use of MPEG2 transport stream.

#### 2.1. MPEG 2 Transport Stream

Despite the different encoding formats which have gradually appeared in the market, the most extended video transmission format for the contribution, distribution and broadcasting of professional quality video signals continues to be the MPEG2 standard [1].

In the said standard two types of formats are specified, the transport stream and the program stream. The first is used for transmission due to its greater robustness concerning noises in the channel and the second is used for production in environments with low error rates.



The various errors which may occur during transmission of the transport stream are corrected at reception, so as to minimize the effect that these may produce in the image. Multiple methods have been developed for this purpose.

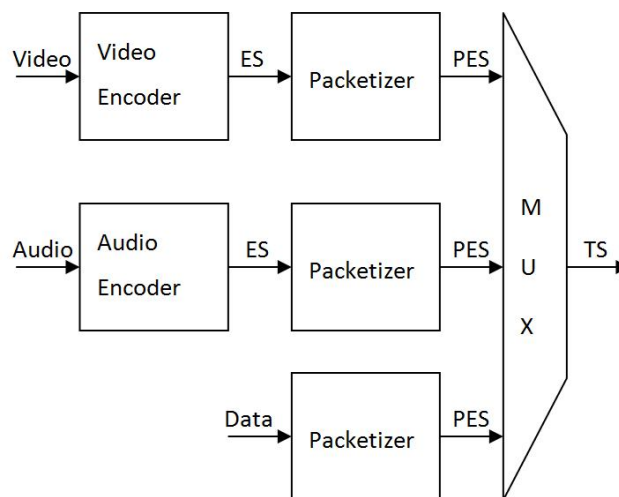


Figure 1. A program multiplexed on a transport stream.

A program comprises several types of data (video, audio, and data) which are encapsulated into elementary streams (ES) and multiplexed in a data stream. Each elementary stream is packaged into Packetized Elementary Stream (PES packet). In order to maintain synchronization between the audio and video data, time stamps are inserted for a correct decoding and displaying of images and sound.

Fig. 1 shows the multiplexing of a video signal, audio signal and other data associated with a program stream as in [4]. The speed of the elementary stream may vary depending on the quality required for the images. In the distribution to broadcaster, the speed may vary from 8 to 50 Mbps. The nature of the images and the transmission purpose will determine the selected quality. For signal broadcasting, 2.5 to 7 Mbps are generally used.

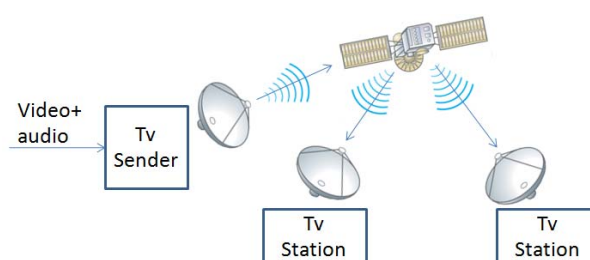


Figure 2. TV distribution.

For the distribution of monitoring channels, either DVB or VoIP, the channels can be compressed to higher rates (i.e. from 4 to 8 Mbps) and several programs can be multiplexed as a single transport stream.

In either case, previous data is accessible from the transport stream which is, thus, generated at the source with the application of inverse operations from the transport stream.

Fig. 2 shows the signal distribution to the different broadcasters.

## 2.2. File Transfer

Signal distributions are generally performed through dedicated link. This type of video connections is data unidirectional links in charge of transmitting the transport stream. At the reception, it is installed the corresponding demultiplexer.

For file distribution over unidirectional links, there are various file transfer protocols based on retransmission patterns of the same file. The Reliable Multicast Transport (RMT) IETF Working Group deals with the standardization of reliable one-to-many multicast transport protocols.

In [5], a study discusses three types of transfer protocols which can be used in unidirectional networks. The Asynchronous Layered Coding (ALC)[9] does not require any type of feedback from the receivers, and the data are encoded using FEC codes. Repetitions of symbol transfer guarantee the integrity of the file at the expense of diminished effectiveness in the bandwidth.

The Nack Oriented Reliable Multicast (NORM) [10] retransmits only the damaged parts from some of the receptors which send signals of Negative Acknowledgments (NACK) over damaged blocks.

The File Delivery over Unidirectional Transport (FLUTE)[11] [12], based on the ALC protocol, with the extension to be used in any type of transmission channel (unidirectional or not) presents metadata which complete the image signal itself (e.g. File name, codec, etc.). Examples for one-way services can see in [13],[14] y [15].

## 2.3. Distribution Format Exchange Profile (DFXP)

Actually EBU has adopted the Timed Text (TT) Authoring Format 1.0 – Distribution Format Exchange Profile (DFXP) and has created a working group for use the standard DFXP for exchange subtitling information in XML.

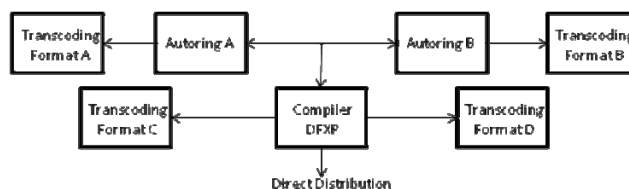


Figure 3. System Model, timed text authoring.

The timed text authoring format is a content type that represents timed text media for the purpose of interchange among authoring systems. Timed text is textual information that is associated with timing information, it serves as a bidirectional interchange format among a heterogeneous collection of authoring systems, and as a unidirectional interchange format to a heterogeneous collection of distribution formats after transcoding or compilation to the target distribution formats as required, and where one particular distribution format is DFXP. Authoring users produces, exchange data, transcode information to different formats and compile to DFXP for distribution to DFXP clients or transcoding to other formats as see in Fig. 2.

A simple example of content is:

```

<body region="subtitleArea">
  <div>
    <p xml:id="subtitle1" begin="0.16s" end="3.15s">
      Kaboom </p>
    </div>
</body>

```

Where a subtitle "Kboom" is presented in the image between seconds 16" and 3'15" when the picture file associated with the subtitle are played. The standard provides more fields in order to indicate other characteristics about the subtitle (position, color, font, etc.).

### 3. Proposed System

#### 3.1 Workflow

Subtitle live events are usually done both with speech recognition and/or special keyboards, data are sending directly to the playout system in order to add the subtitle data to MPEG2 stream.

New workflow is based in two phases: file and identifier distribution.

1) File distribution. Before the data transmission event producer's send a XML file according with DFXP metamodel to the entire broadcaster involved in the transmission, the file can be transmitted in conventional way.

Broadcaster translates the information to desired languages, original document can be one of output set. In Fig. 4 we can see the workflow of file from producer to broadcaster in order to translate the text in the desired languages before the transmission date.

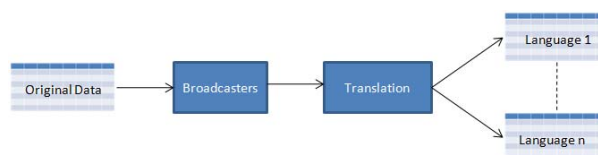


Figure 4. File Distribution workflow.

2) Identification distribution. Identification of each subtitling item is sent by the event producer company as data PES packet inside the MPEG2 transport steam through video link. Each subtitle PES packet can be retransmit several times with FLUTE or NACK protocol in order to avoiding errors in the transmission channel. This protocol has good response with few receptors number [5]. High error rates can cause the impossible to know the identification; the reception can send a NACK to transmitter in order to retransmit the packet.

Live events have to work in real time, if audio are delayed more than 200 milliseconds from video the spectators can note the delay and the results it's very worry but for subtitle case this restrictions is less strong than audio channel, subtitles can be delayed one second maintaining the exposure time.

When the valid identification data is arrived in the TV station, the identification is searching in the language documents for send the appropriated contents depending on the diffusion transmission.

In analogue video transmission, ASCII characters are included in the vertical retrace interval right before the first visible horizontal line. Simple decode circuitry was mandated to be included in all TVs that would provide the extraction and storage of the data and allow the TV user to add the closed caption characters as an overlay to the video in the next fields.

The advent of MPEG compressed video brought with it more possibilities for higher bandwidth closed captioning channel and other supplementary data channels. The European Telecommunications Standards Institute (ETSI) specifies the method by which subtitles, logos and other graphical elements may be coded and carried in DVB bitstreams [16].

Fig.5 shows haw Producer Company adds data packets to video channel. Broadcaster extracts the subtitle identifier from transport stream and selects the associated text with its characteristics and can insert the

information in the video signal and/or sent as data text several language channels for visualize at user demands. Subtitle information is adapted for each diffusion technologies.

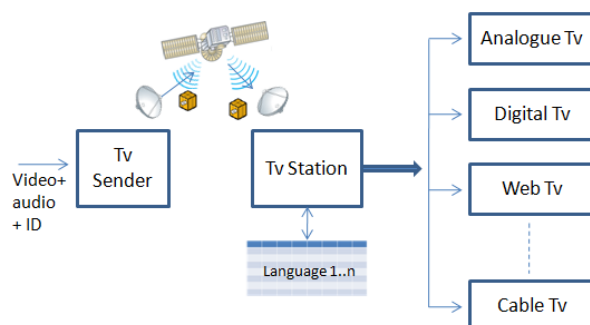


Figure 5. Identification transmission.

### 3.2 Data Model

Timed Text file has included all the necessary data for subtitle except the time in which the subtitle will be displayed and its duration, absolutely necessary for stored programs. The exact moment of subtitle displaying is unknown in live events. It is necessary to complete this information in real time. The text to be displayed is within a "P" element in the DFXP metamodel. A "P" element represents a logical paragraph, serving as a transition between block level and inline level formatting semantics and it has his corresponding identification attribute. This attribute will be used to link original subtitle format and text to the corresponding translated text, as explained above. There are also two important attributes: begin and duration.

Begin and duration attributes are time expressions that can be a clock time or an offset. The span of time a subtitle is about to be displayed is included in the original file as offset time expressions, setting the begin value to an estimate time and the duration attribute to the corresponding offset (usually, only few seconds more). When transmission is in course, begin value will be set in real time according to the show needs, (for example, applause delays or unexpected events). The begin value will be taken from a counter that starts with the show. The XML structure of a "P" element is shown in figure 6.

The following example shows a processed subtitle to be included, and the next subtitle to be processed:

```

<p xml:id="subtitle98" begin="32m0.76s" duration="1.45s">
  Hello Figaro,
</p>
  
```

Note that "subtitle98" is displayed at 32 minutes from the beginning of the show and "subtitle99" is waiting for an appropriate begin value. In this example, "subtitle99" begin estimate time is "30m8s", that is, the show is delayed two minutes from the estimate timing.

```

<p xml:id="subtitle99" begin="30m8s" end="2.0s">
  Are you alone?
</p>
  
```

The second subtitle will be displayed during two seconds, but the new begin value will be set in real time when it need to be displayed.

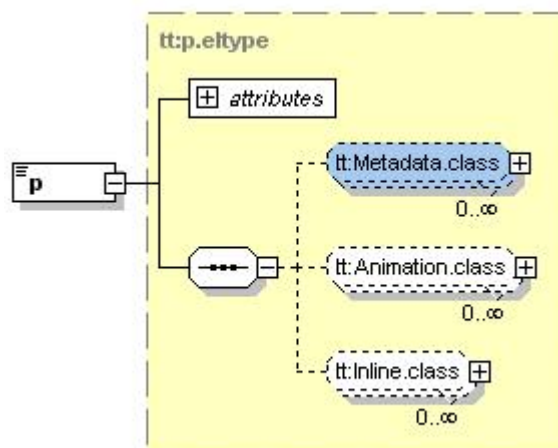


Figure 6. XML structure of a "P" element.

#### 4. Conclusion

TV production and broadcasting model are changing with the introduction of new technologies in distribution processes; the growing of new digital TV channels and the new cheaper communications networks facilitates to share the production between several broadcasters companies.

Event producer companies have the knowledge, broadcaster only retransmitting the video and audio signal with the playout system but usually they don't add event information.

Data inclusion inside the video signal help the knowledge transmission between event producers companies to end user while broadcasters only have to adapt video, audio and data content to the different diffusion technologies.

With this work we open an initiative for create a new production services provided from Producers to Broadcaster, this service will help to end user to understand Opera without cost increase.

#### Bibliography

- [1] W3C Timed Text (TT) Authoring Format 1.0 – Distribution Format Exchange Profile (DFXP). Candidate Recommendation 16 November 2006. Available in: <http://www.w3.org/TR/ttaf1-dfxp/>
- [2] SMPTE 377M-2004. Television — Material Exchange Format (MXF) — File Format Specification
- [3] ISO/IEC 13818-1, Information Technology-Generic Coding of Moving Pictures and Associated Audio Information, Part 1: System, April 1996.
- [4] Stephan W. Mondwurf, Low Cost MPEG-2 Multiplexing Scheme for Multimedia and Digital TV Distribution Systems, Proceedings of the Fifth IEEE International Caracas Conference on Devices, Circuits and Systems, Dominican Republic, Nov.3-5, 2004.
- [5] Neumann, C., Roca, and V. Walsh, R. 2005. Large scale content distribution protocols. SIGCOMM Comput. Commun. Rev. 35, 5 (Oct. 2005).
- [6] M. Luby, J. Gemmell, L. Vicisano, L. Rizzo, and J. Crowcroft. Asynchronous Layered Coding (ALC) protocol instantiation, Dec. 2002. Request For Comments 3450.
- [7] B. Adamson, C. Bormann, M. Handley, and J. Macker. Negative-acknowledgment (NACK)-Oriented Reliable Multicast (NORM) Protocol, Nov. 2004. Request For Comments 3940.

- [8] T. Paila, M. Luby, R. Lehtonen, V. Roca, and R. Walsh. FLUTE - File Delivery over Unidirectional Transport, Oct. 2004. Request For Comments 3926.
- [9] U. Reimers, DVB—The Family of International Standards for Digital Video Broadcasting, Proc. IEEE, vol. 94, no. 1, pp. 173–181, Jan. 2006.
- [10] U. Reimers, DVB (Digital Video Broadcasting), Springer Verlag Berlin, 2nd Edition 2004.
- [11] Bürklin, H., Schäfer, R., and Westerkamp, D. 2007. DVB: from broadcasting to ip delivery. SIGCOMM Comput. Commun. Rev. 37, 1 (Jan. 2007), 65-67.
- [12] T. Paila, M. Luby, R. Lehtonen, V. Roca, and R. Walsh. FLUTE - File Delivery over Unidirectional Transport, Oct. 2004. Request For Comments 3926.
- [13] E. Pallis, C. Mantakas, G. Mastorakis, A. Kourtis, and V. Zacharopoulos, Digital Switchover in UHF: the ATHENA concept for broadband access, European Transactions on Telecommunications, 17, p.175–182, 2006.
- [14] MOBISERVE EU IST project, New mobile services at big events using DVB-H broadcast and wireless networks, <http://www.mobiserve.org/>.
- [15] R. Schatz, S. Wagner, and N. Jordan, Mobile Social TV: Extending DVB-H Services with P2P-Interaction, International conference Digital Telecommunications (ICDT), 2007.
- [16] European Telecommunications Standards Institute. 1997 ETS 300 743 Digital Video Broadcasting (DVB); Subtitling systems.
- 

### Authors' Information

---

*Jesús Martínez Barbero – OEI, E. Universitaria de Informática; Universidad Politécnica; Madrid, Spain;*  
*e-mail: [jmartinez@eui.upm.es](mailto:jmartinez@eui.upm.es)*

*Manuel Bollain Pérez – OEI, E. Universitaria de Informática; Universidad Politécnica; Madrid, Spain;*  
*e-mail: [mbollain@eui.upm.es](mailto:mbollain@eui.upm.es)*

---

## PERFORMANCE ANALYSIS OF CALL ADMISSION CONTROL FOR STREAMING TRAFFIC WITH ACTIVITY DETECTION FUNCTION

Kiril Kassev, Yakim Mihov, Boris Tsankov

*Abstract:* Admission control is a key issue for quality of service (QoS) provisioning in both wired and wireless communication networks. The call admission control (CAC) algorithm needs to know the source traffic characteristics and the required performance in order to determine whether the connection can be accepted or not and, if accepted, the amount of network resources to allocate. In this paper, we determine the CAC threshold value in case streaming homogeneous ON-OFF traffic flow is considered. An analytical method for packet loss probability evaluation is proposed and numerical examples are presented.

*Keywords:* streaming traffic, ON-OFF traffic model, call admission control, packet loss probability

*ACM Classification Keywords:* C.2.1 Network architecture and Design – Wireless communication, C.2.5 Local and Wide-Area Networks – Access schemes

---

### Introduction

The necessity of call admission control (CAC) arises together with the wide deployment of connection-oriented packet switching technologies. CAC is the name for a set of tools which has to take a decision whether or not a new connection can be served by the system, in addition to those of the connections that are in progress. If the new connection is admitted it will not deteriorate the bandwidth usage and performance of the connections already established. CAC is a fundamental mechanism for congestion control and QoS provisioning. For this reason, it has been extensively studied in both wired [1] and wireless [2] networks.

The CAC design and performance analysis became an inseparable part of ATM and IP based networks planning [3] and different wireless networks as well. CAC mechanisms can be classified based on various objectives and design options. Among the aims one can list: QoS parameters like call level and packet level congestion probabilities, packet delay, bandwidth guarantee; Optimization of throughput, power allocation, fairness; Controlling handover failure probability, etc. In this paper, the *call blocking* and the *packet dropping* probabilities are considered.

According to the decision time, CAC schemes can be classified as proactive (parameter-based) and reactive (measurements-based). In the former scheme, the arriving call is permitted or rejected based on predictive analytical evaluation of the QoS constraints. In the latter scheme, the CAC decides to permit or reject the call dynamically based on some QoS measurements. Both approaches have advantages and disadvantages. A combination of these two approaches could be used for more effective congestion control to be provided. In this paper, the *proactive CAC* is considered.

The subject of our interest is the traffic flow generated by multiple *variable bit rate* sources and the *bursty traffic* in particular where each source is represented as an ON-OFF source. Our considerations are restricted to *streaming (real-time) traffic* generated by VoIP sources and other multimedia sources as well. There are two consequences of this:

- a) It is not possible to compensate the *burst-scale losses* by means of a buffer due to the very stringent restrictions on packet delay and the long average burst duration  $T_{ON}$  ;
- b) It is reasonable to apply the so called *bufferless fluid flow* or *burst-scale loss approach* ([3], Chapter 12). The buffer (or buffers) used is relatively small and it is dimensioned to cope with the packet-scale losses only.

In this paper the bufferless fluid flow approach is used for CAC parameters determination. The assumption that there is not a buffer at burst level leads to a conservative estimates for packet losses and therefore to the safety side of CAC parameter determination.

---

### The Traffic Model

---

The bufferless fluid flow model is used quite a while ago [4] – [7] due to its effectiveness and simplicity. Our considerations are restricted to homogeneous traffic sources with the most popular example – the VoIP traffic. Due to implementation of a voice activity detection algorithm into a voice codec of a particular type (i.e. G.729), an *ON-OFF traffic source* is usually characterized by means of the following three parameters: the bit rate during a burst (burst rate)  $R$ ; the mean bit rate  $r$  and the burst duration  $T_{ON}$  . From the obvious relation

$$r = \frac{T_{ON}}{T_{ON} + T_{OFF}} R \quad (1)$$

one can obtain  $T_{OFF}$  . In the same time it is evident that an ON-OFF source could be characterized by means of any three out of the four parameters in (1).

Let us suppose  $C$  represents the total bit rate allocated to streaming traffic transmission, for example UL or DL radio-link capacity of an IEEE WiMAX or 3GPP LTE (Long Term Evolution) base station. We also define the following notations:

- $n$  is the maximum number of active traffic sources that can be simultaneous served (or number of transmission resource units);
- $N$  is the maximum number of calls (sessions) admitted to the system by the CAC.

The value of  $n$  is expressed as

$$n = \frac{C}{R} \quad (2)$$

Our aim is to determine  $N$  as a quantity of CAC. Obviously, there is no sense the value of  $N$  to be less than  $n$ . If  $N = n$ , all admitted calls will be served without any packet losses. In order to utilize the ON-OFF traffic behavior due to activity detection function,  $N$  has to be more than  $n$ . But if  $N > C/r$ , and the number of admitted calls approaches  $N$ , the buffer will permanently overflow. Therefore, apparently there is  $n < N < C/r$ . The exact value of  $N$  is the maximum one for which the probability  $P_{PL}$ , given packet to be lost is below a prescribed limit.

The system can be in any state  $(i, j)$ , where  $i$  ( $i=0, 1, \dots, N$ ) is the number of accepted calls and  $j$  ( $j=0, 1, \dots, i$ ) is the number of active calls (number of bursts in progress). The call flow forms a Poisson process with call rate  $\Lambda_c$  and call service time  $1/\mu_c$ , whereas the burst flow forms a Binomial process with single OFF source burst arrival rate  $\lambda_b = 1/T_{OFF}$  and single ON source burst service rate  $\mu_b = 1/T_{ON}$ . The combination of both processes forms the state-transition diagram shown on Fig. 1.



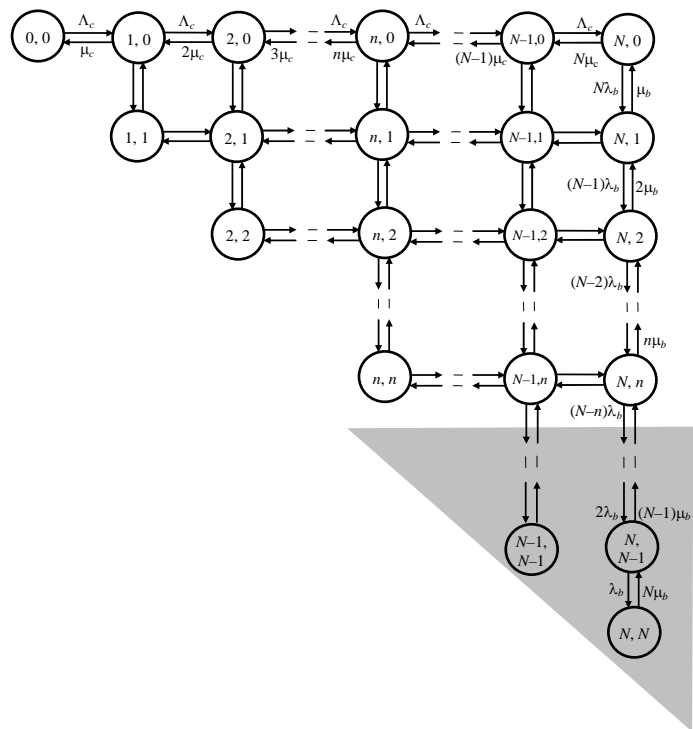


Fig. 1 Two-dimensional state-transition diagram

In the majority of cases when we determine  $N$  there is a common practice of considering the case  $i = N$  only and evaluate  $P_{PL}$  ([3], p. 141). This corresponds to the case where all traffic sources will suffer highest losses and it is related to the most right column on state-transition diagram (Fig. 1). Since we are interested in packet loss probability evaluation, applied for a more realistic case, it is of significant importance to take into account all possible states and perform an analytical evaluation of  $P_{PL}$ .

### Analytical Evaluation

The state-transition diagram on Fig. 1 presents a two-dimensional *Markov process* where the burst level offered traffic depends on the established call number  $i$ .

According to the Erlang-B formula, the probability of exactly  $i$  sources being busy is

$$P(i) = \frac{A_c^i}{i! \sum_{x=0}^N \frac{A_c^x}{x!}} \tag{3}$$

where  $A_c = \Lambda_c / \mu_c$ .

The conditional probability of  $j$  sources being active given that  $i$  traffic sources are busy is

$$P(j/i) = \frac{i!}{j!(i-j)!} \alpha^j (1-\alpha)^{i-j} \tag{4}$$

where

$$\alpha = \frac{T_{ON}}{T_{ON} + T_{OFF}} \tag{5}$$

The overall probability  $P(i, j)$  is

$$P(i, j) = P(i).P(j/i). \quad (6)$$

The offered rate in state  $(i, j)$  is  $j.R$ . The excess rate in the same state  $(i, j)$  is  $(j - n).R$ . As a consequence, the excess rate mean value is given by

$$\sum_{i=\lceil n \rceil}^N \sum_{j=\lceil n \rceil}^i R(j-n)P(i, j) \quad (7)$$

We should note that value of  $n$  is not necessary to be an integer, and  $\lceil n \rceil$  denotes the minimum integer value greater or equal to  $n$ .

The packet loss probability is given by the relation

$$P_{PL} = \frac{\sum_{i=\lceil n \rceil}^N \sum_{j=\lceil n \rceil}^i R(j-n)P(i, j)}{\sum_{i=1}^N \sum_{j=1}^i RjP(i, j)} \quad (8)$$

The equation (8) can be simplified, and hence

$$\begin{aligned} P_{PL} &= \frac{\sum_{i=\lceil n \rceil}^N \sum_{j=\lceil n \rceil}^i (j-n)P(i, j)}{\sum_{i=1}^N P(i) \sum_{j=1}^i jP(j/i)} \\ &= \frac{\sum_{i=\lceil n \rceil}^N \sum_{j=\lceil n \rceil}^i (j-n)P(i, j)}{\sum_{i=1}^N P(i)i\alpha} \end{aligned} \quad (9)$$

In order to perform a comparative analysis, considering the case where  $i = N$  ([3], p.141)  $P_{PL}$  could be derived from (8)

$$P_{PL} = \frac{\sum_{j=\lceil n \rceil}^N (j-n)P(j/N)}{N\alpha} \quad (10)$$

---

## Numerical Results

---

The current section deals with performance evaluation of analytical model proposed. In order to decrease the bandwidth usage the encoding scheme of each traffic source employs an activity detection function, which is quantitative represented by the activity factor  $\alpha$ . Thus, the offered traffic flow  $A_c$  is generated by multiple homogeneous ON-OFF sources. Due to the limited amount of system resources available, the maximum number of calls (sessions) admitted to the system depends on the target call (session) blocking probability  $B$ , which can be obtained by (3).

Based on the required performance thresholds, such as  $B$  and  $P_{PL}$ , as well as source traffic characteristics (i.e.  $A_c$ ), the significant task of CAC is to determine whether the connection can be accepted or not and, if accepted, the amount of network resources to be allocated. Fig. 2 shows the comparison results of the network dimensioning with typical values of the packet losses  $P_{PL}$  and activity factor [8], by applying both the model presented in [3] (10) (we will refer to it as "model A") and proposed analytical model (9) (we will refer to it as "model B"). It should be noted that "model A" refers to a case when the system is heavy loaded ( $i = N$ ). Since the network service providers are interested in the system performance evaluation under normal load condition, it is necessary all possible system states to be taken into consideration. Numerical results show that this led to more

efficient network resource usage (bandwidth), since the system is not overdimensioned, as it is done by using (10). On the other hand, as expected, silence suppression considerably decreases the transmission resource usage needed to meet the target packet loss probability.

According to [8], the degradation of voice quality (subjective quality measure MOS) is tightly coupled with the packet losses and coding scheme used. In case of using G.729, the admission of 2 % packet losses reduces the MOS from 4.0 to 2.75. Fig. 3 shows network dimensioning for different values of  $P_{PL}$ . Study results demonstrate that the same amount of network resource could be allocated to meet call flow demands with higher value of activity factor ( $\alpha = 0.6$ ), compared to the case when  $\alpha = 0.45$ , but there is a trade-off that users will experience poor voice quality (MOS is reduced to a value of 2.75 [8]). This could be applied in case of short-term resource reduction.

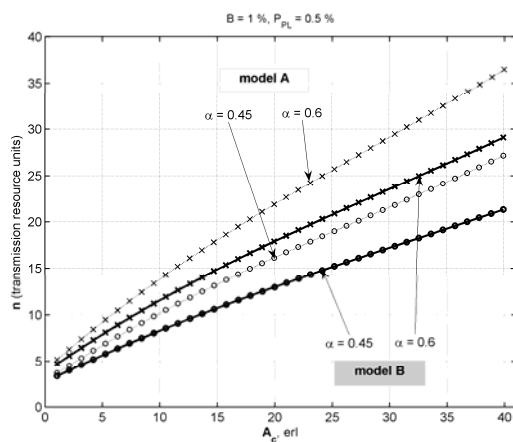


Fig. 2 Network dimensioning – CAC models comparison

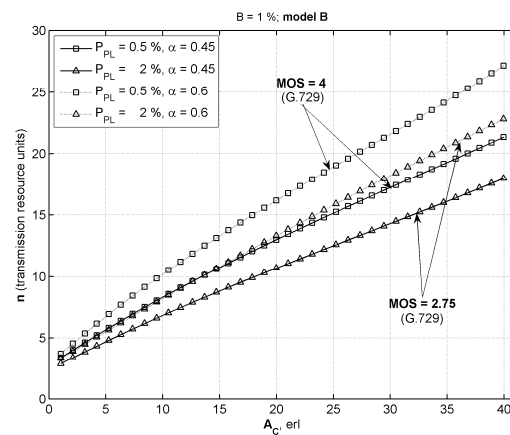


Fig. 3 Network dimensioning – model B

## Conclusion

In this paper, an analytical method for quantitative analysis of a call admission control mechanism is proposed. The method takes into account the more realistic case study of bursty traffic arrival. A comparative analysis with a model with similar capabilities, suggested in the literature, has been performed. The results obtained demonstrate that analytical method proposed is efficient, especially when it is applied for wireless access networks dimension, since it does not overdimension the network in terms of necessary bandwidth (transmission resource units). This is of significant importance, since wireless communications resources are scarce and expensive. The method developed can be used in cross layer design of wireless networks, where considering the application requirement the more efficient resource allocation is achieved [9]. It is based on the Erlang model on the call level although an assumption of the Engset model [10] is also applicable with corresponding numerical complexity. The model could be extended by considering a heterogeneous traffic case.

## Acknowledgement

This paper is a part of research work in the context of the research project "Optimal telecommunication resource allocation considering cross-layer interaction", funded by the Bulgarian Ministry of Education and Science with grant DVU01/0109, as well as a research project funded by the Research and Development Sector of the Technical University of Sofia with grant 080910dni-7.

---

## Bibliography

---

- [1] H. G. Perros and K. M. Elsayed, "Call admission control schemes: a review", *IEEE Communications Magazine*, vol. 34, no. 11, Nov. 1996, pp. 82-91.
- [2] M. H. Ahmed, "Call admission control in wireless networks: A comprehensive survey", *IEEE Communications Surveys*, vol. 7, no. 1, 2005, pp. 50 – 69.
- [3] J. M. Pitts and J. A. Schormans, *Introduction to IP and ATM Design and Performance* (Second edition), Chichester, England: John Wiley & Sons, 2000.
- [4] J. Y. Hui, "Resource allocation for broadband networks", *IEEE J. on Selected Areas in Communications*, vol. 6, no. 9, Dec. 1988, pp. 1598-1608.
- [5] R. J. Gibbens, F. P. Kelly and P. B. Key, "A decision-theoretic approach to call admission control in ATM networks", *IEEE J. on Selected Areas in Communications*, vol. 13, no. 6, Aug. 1995, pp. 1101-1114.
- [6] M. Reisslein, K. W. Ross and Rajagopal, "Guaranteeng statistical QoS to regulated traffic: the single node case", *Proc. 18th, IEEE INFOCOM*, 1999, pp. 1061-1072.
- [7] G. Mao and D. Habibi, "Loss performance analysis for heterogeneous ON-OFF sources with application to Connection Admission Control", *IEEE/ACM Trans. on Networking*, vol. 10, no. 1, Feb. 2002, pp. 125-138.
- [8] A. P. Makropoulou, F. A. Tobagi, M. J. Karam, "Assessing the quality of voice communications over Internet backbone", *IEEE/ACM Transactions on Networking*, Oct 2003, vol. 3, no. 5, pp. 747-760.
- [9] E. Pencheva, I. Atanasov and D. Marinska, "Cross Layer Design of Application-level Resource Management Interfaces", *Proc. of IEEE International Workshop on Cross Layer Design IWCLD'2009*, June 2009, Spain, pp. 1-5.
- [10] V. G. Vassilakis, G. A. Kallos, I. D. Moscholios and M. D. Logothetis, "The wireless Engset multi-rate loss model for the call-level analysis of W-CDMA networks", *IEEE 18th Int. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2007)*, 2007, pp. 1-5.

---

## Authors' Information

---

*Kiril Kassev* – Assistant Professor, Department of Communication Networks, Technical University of Sofia, 8 Kliment Ohridski Blvd., Sofia-1000, Bulgaria; e-mail: [kmk@tu-sofia.bg](mailto:kmk@tu-sofia.bg)

*Yakim Mihov* – MSc Student, Department of Communication Networks, Technical University of Sofia, 8 Kliment Ohridski Blvd., Sofia-1000, Bulgaria; e-mail: [yakim\\_mihov@abv.bg](mailto:yakim_mihov@abv.bg)

*Boris Tsankov* – Professor, Department of Communication Networks, Technical University of Sofia, 8 Kliment Ohridski Blvd., Sofia-1000, Bulgaria; e-mail: [bpt@tu-sofia.bg](mailto:bpt@tu-sofia.bg)

---

## ANALYSIS OF MALICIOUS ATTACKS ACCOMPLISHED IN REAL AND VIRTUAL ENVIRONMENT

Dimitrina Polimirova, Eugene Nickolov

*Abstract:* In this paper an analysis of possibilities offered by virtual environments for accomplishing attacks to and within it, is made. Main techniques for accomplishing an attack to virtual environment are pointed and real virtual attacks and successful virtual attacks are examined. An analysis of accomplished attacks is made and percent distribution of real virtual attacks and successful virtual attack is graphically illustrated. Respective assessments and recommendations for future investigation are made.

*Keywords:* Virtual Machine Environment, Virtual Environment, Attack/Attack tools, Defense/Defense tools, Malicious Code

*ACM Classification Keywords:* D.4.6 Security and Protection: Invasive software (e.g., viruses, worms, Trojan horses)

---

### Introduction

Computer Viruses, Worms, Trojan Horses, Backdoor, Rootkits, Spyware, Adware, etc. are terms used years ago mostly by computer specialists. However, now they present in daily speech not only of employments of corporate, academic and government organizations, but also of final users. Everybody knows less or more for their action and damages they can harm to the computer systems and information flows. The availability of various attack methods determines the necessity of investigation of different methods and means for defense of computer systems, networks and information flows.

A general strategy for protecting computer systems and networks could include using virtualization techniques to achieve increase in information security not only of information flows, represented by file objects, but also of operation system as a whole.

Since the 70<sup>th</sup> of XX century the problem for security and protection of information flows has drawn developers' and constructors' attention in the area of information technology [1]. With the first malicious attack in the 60<sup>th</sup> of last century [2], investigations in the area of system protection receive financial support. As a result for a short time ideas decreasing the risk in the system management, are realized.

---

### The aim and tasks

The aim, which can be set in this paper, is related to investigations of the set of possible virtual attacks and its reduction to the set of successful virtual attacks, which have specific behavior in Virtual Machine Environments (VMEs) to overcome its protecting mechanisms.

The main hypothesis will be related to the possibility to make analysis and estimation of different techniques for successful accomplishing of an attack to virtual environment.

The following tasks are set in reaching the aim:

- 1) to make analysis of Virtual Machine Environment as a possibility for accomplishing attacks to and within it;

- 2) to determine the set of real virtual attacks;
- 3) to examine the main techniques for accomplishing an attack to virtual environment;
- 4) to determine the set of successful virtual attacks.

---

## The Problem

---

### 1. VIRTUAL MACHINE ENVIRONMENT AS A PLACE FOR EXECUTING MALWARE

#### 1.1. Description of Virtual Machine Environment

Virtual Machine Environment (VME), mention also in this paper as virtual environment, gives the possibility to create one or more guest operating systems from one primary (host) operating system. Each created guest operating system works in emulated environment and it has controlled access to virtual and real hardware resources (Figure 1).

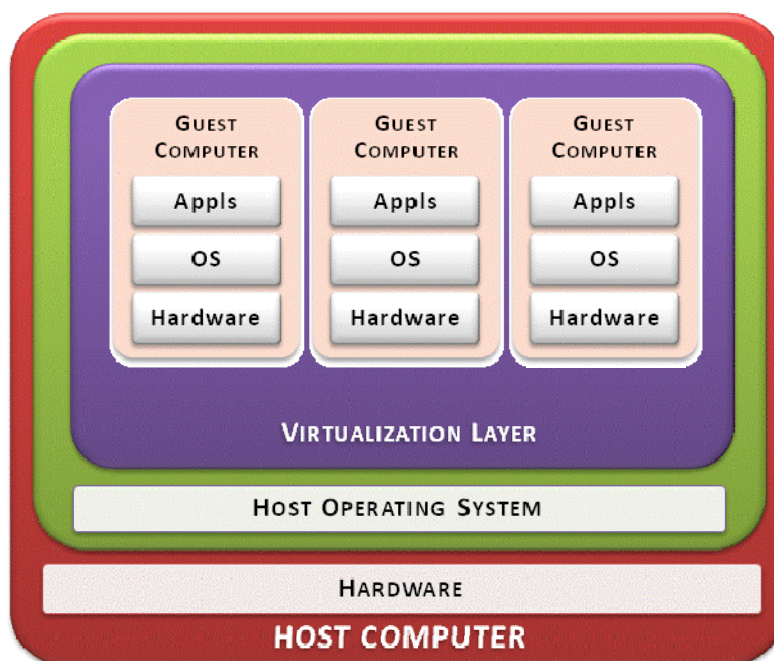


Figure 1 Virtual Machine Environment

#### 1.2. Benefits of Virtual Machine Environment.

The increasingly using of virtual environments speaks for itself for the benefits, which can be obtained, namely:

- (1) possibility for reducing the number of physical machines. This can be achieved by the possibility, which virtualization offers to integrate several servers into one hardware platform. This way expenditure for hardware can be drawn;
- (2) easy management. The management is from one physical place by one person. We can see here the economic benefit with respect to the human resources, but more important is that one person, which is

specialist has an access to the management of real and all virtual environments. This way the possibility to harm the system by someone through carelessness or through ignorance is decreased significantly;

- (3) increasing the security. The virtual environments are exposed to attack as real environments. Since the virtual environments are identical to real one, they are friendly place for distributing of malicious software and accomplishing of attacks. Not every employee in one organization has enough experience and knowledge to succeed to protect itself from such types of attacks even though most of the employees are more or less familiar to the main actions for protecting of malicious software. From the point of view of the primary operating system, the virtual environment can have a defense role. All events occurred in the virtual environment are separated from the real environment. In the most cases if the malicious software gets into the guest operating system, it continue to thrive there while the primary operation system, host computer and even other virtual machines stay clean;
- (4) possibility to load different operation systems on one hardware platform. This technique is often used by computer specialists in the processes of building and testing different software. Different computer system and networks configurations are simulated. If there is a program bug, the host operation system leaves unaffected;
- (5) possibility for easy and fast recovery of critical applications. In case of system crash the virtual environment gives the possibility for fast and easy recovery not only of critical applications, but also of the whole system.

### 1.3. Virtual environment in the practice

In most cases virtual environments are used by software vendors during the processes of software building and testing. This technology can take place in the work of different sphere, of course - individual users, businesses, government agencies, and academic institutions. For the aim of this paper only the virtual environments application in the work of security and defense vendors on the one hand, and malicious code vendors on the other hand, will be examined.

Using the virtual environments became so popular that hackers direct their attention to them. Attacks, accomplished to virtual environment, are investigated not only by the vendors of security and defense software, but also by the vendors of malicious codes.

Vendors of antivirus and security software use virtual environments, to examine the behavior of different types of attack tools without damaging the real environment. After that the respective defense tool can be built.

On the other hand vendors of malicious code use virtual environments too to test if their attack tool works appropriately and can accomplish its planned action in different computer, system and network configurations.

The virtual environment is that which consolidate the both sides – used by ones to protect, and by others – to attack. In this connection the main goal of the vendors of malicious codes is to detect the presence of virtual environment and if they detect one – to change their action.

### 1.4. Virtual Machine Environments software

Figure 2 shows comparison of the most widespread virtual machine emulators used by hackers.

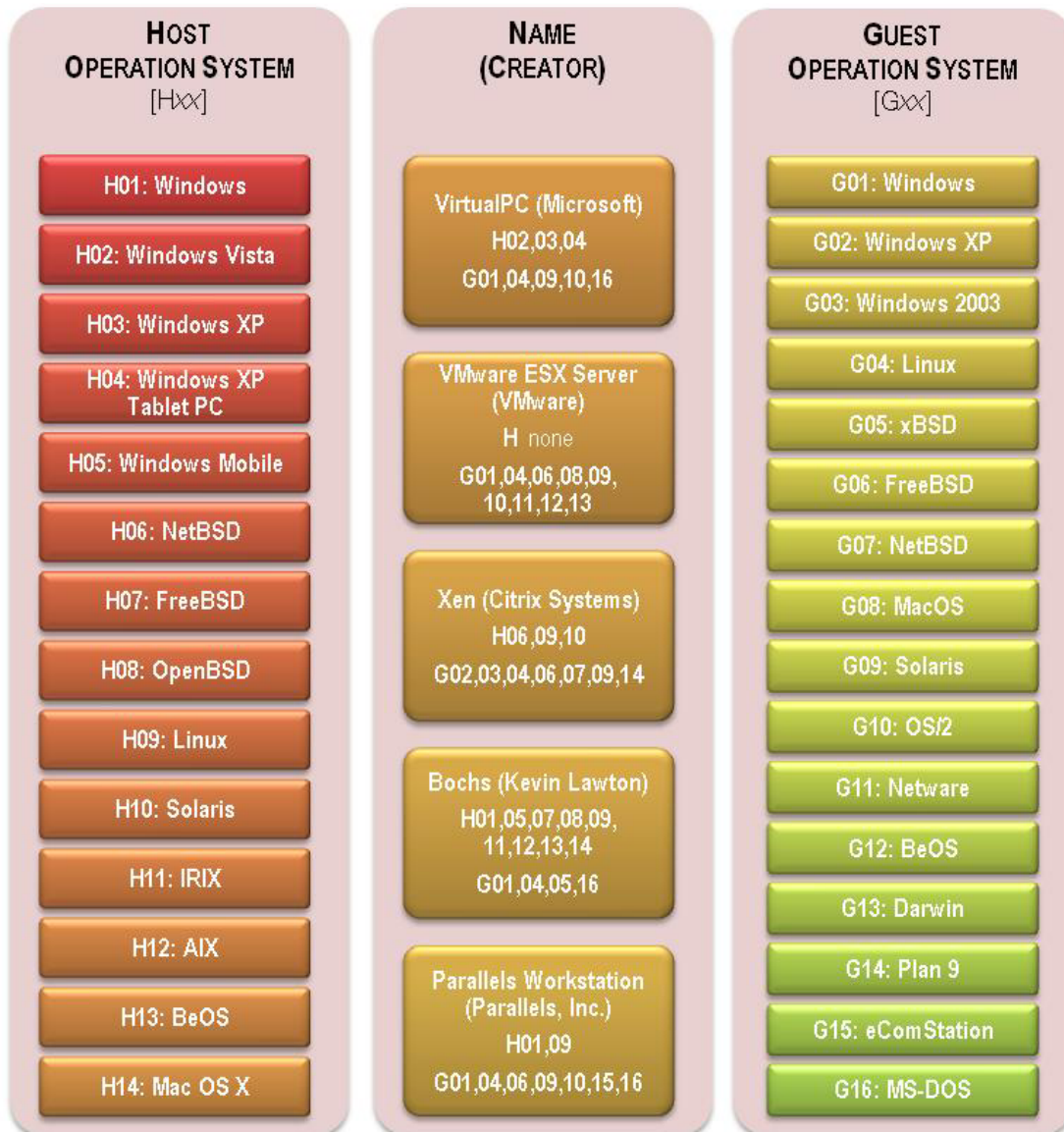


Figure 2 Widespread virtual machine emulators

## 2. SETS OF POSSIBLE VIRTUAL ATTACKS AND SUCCESSFUL VIRTUAL ATTACKS

### 2.1. Analysis of the set of possible virtual attacks

As a whole the attacks can be presented by malicious software and malicious attack. In case of malicious software the direct participation of a user at the moment of the attack is missing, while in case of malicious attack the user's presence is required. [3], [4].

The variety of attack tools for the recent years is big. They can be terminologically divided into two main categories: malicious software (malware) and greyware (grayware). In the set of possible virtual attacks are included attack tools, used successfully or not to accomplish attacks within/to virtual environment. The set of successful virtual attacks includes 11 basic (the most popular for 2008) attack tools (9 for the group of malware and 2 for the group of greyware), separated in 4 groups respectively 3 for malware and 1 for greyware (Figure 3). The chosen attack tools are generalized from the most frequently used attack tools in the recent years [5].



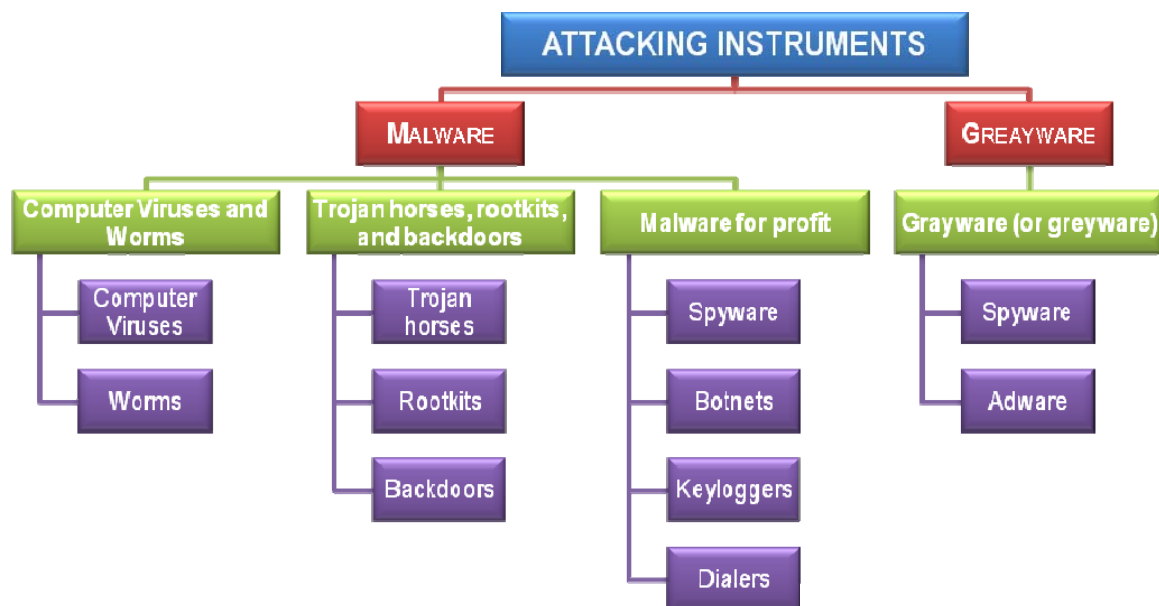


Figure 3 Attack tools included in the set of possible virtual attacks

## 2.2. Main techniques for accomplishing an attack to virtual environment

According to security software vendors the main attacks, accomplished to virtual environment, are three [6]:

- 1) the goal of the first type of attack is to detect the presence of virtual environment;

This is the most widely spread attack to VMEs and it is in pawn in the attack tools' scenarios. The goal of the attacks is to determine the presence of virtual environment. If such is found they change their behavior. In the common case of detecting virtual environment malicious code stops its activity. That way the vendors of antivirus and security software can't easily analyze the malicious code and building the respectively protection tool is hard and more slowly.

Figure 4 shows several techniques for detecting the presence of virtual machine environment [7]. These techniques are used by different attack tools from the groups "Computer Viruses and Worms", "Trojan horses, rootkits, and backdoors", "Malware for profit", and "Grayware (or greyware)" (for example: SubVirt, Confiqer, Vundo, Agent.FDS, Banking.G, OscarBot.UG, Socks, Aresas.a@MM, CodeRed, etc.).

- 2) the second attack consists in the possibility of the malicious code to accomplish DoS attack to VME, which will force the virtual machine to exit;

Usually the attack tool has a vulnerability scan. If the malware finds vulnerability it executes the exploit in Virtual Machine Environments. After that a Denial-Of-Service attack can be accomplished. That technique is used by ByteVerify, MS09-033, etc.

- 3) the third attack consists in the possibility of the malicious code to dismiss the virtual machine protecting environment.

In this case the attack needs to completely dismiss the virtual machine protecting environment and to continue its action. As far as I know such malicious codes don't exist.

Mentioned above malware' names are chosen from the information database of National Laboratory of Computer Virology – BAS as the most frequently accomplished in Bulgaria, Balkan Peninsula and south-east Europe.

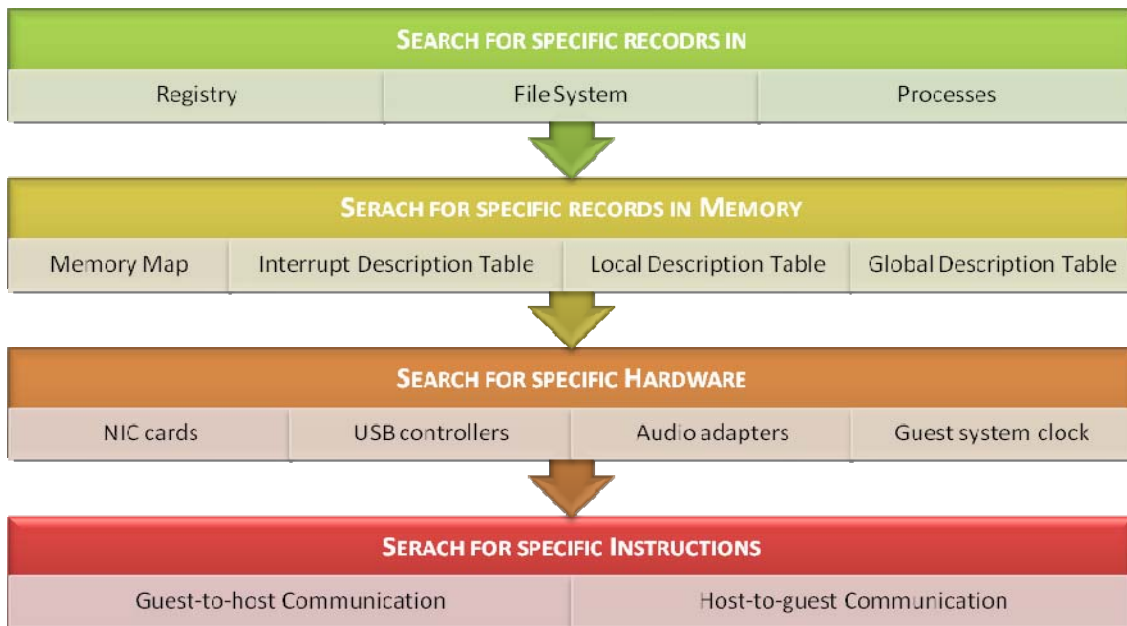


Figure 4 Virtual Machine Environment Detection Techniques

2.3. Analysis of the set of successful virtual attacks

More and more attack tools includes in their code possibility to attack virtual environment. Most tools heaving such a technique are Trojan Horses, Worms, Rootkits, Backdoors, Downloaders and Spyware.

The set of successful virtual attacks includes attack tools that can accomplish an attack to virtual environment. The set of successful virtual attacks includes 7 basic attack tools (6 for the group of malware and 1 for the group of greyware), separated in 4 groups respectively 3 for malware and 1 for greyware (Figure 5).

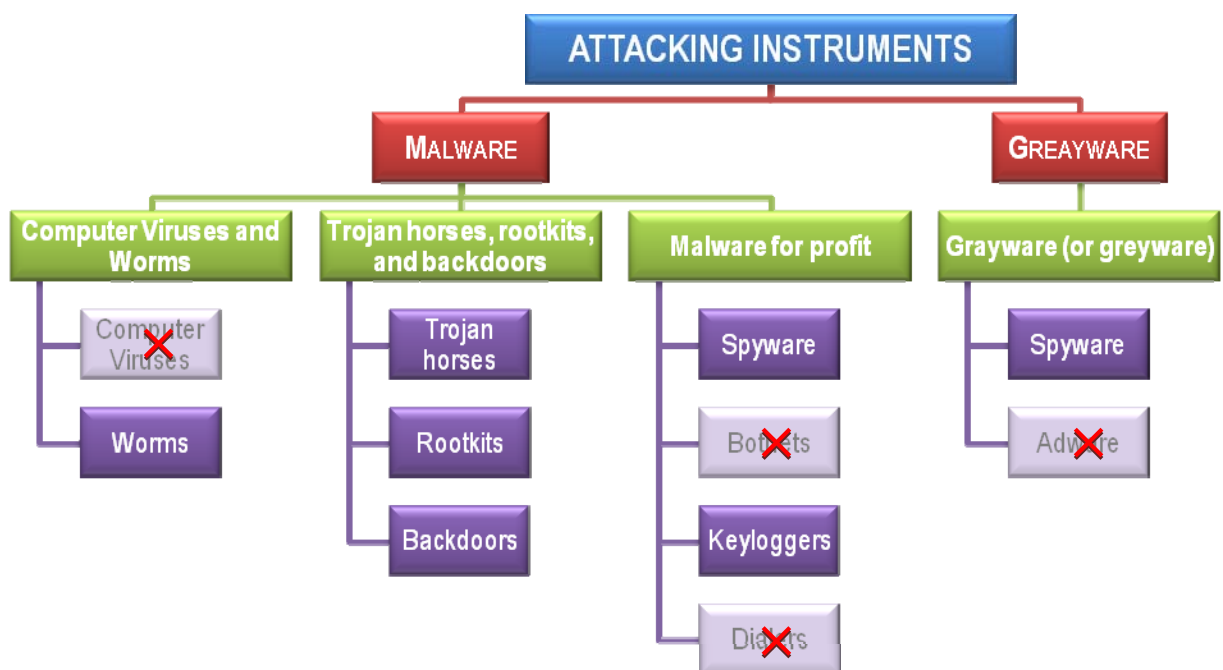


Figure 5 Attack tools included in the set of successful virtual attacks

Figure 6a, b, c, d, e, f shows graphically the percent distribution of the mentioned above attacks, accomplished in Bulgaria, Balkan Peninsula and south-east Europe for 2008. Graphics shows also the percent distribution of the attacks from the set of possible virtual attacks (RE) in relation to respective attacks from the set of successful virtual attacks (VE). The data are collected from the current information base of National Laboratory of Computer Virology of Bulgarian Academic of Sciences.

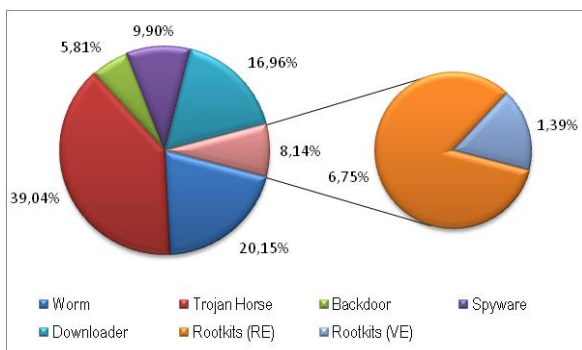


Figure 6a Percent distribution of Rootkits (RE) and Rootkits (VE) Downloader (RE) and Downloader (VE)

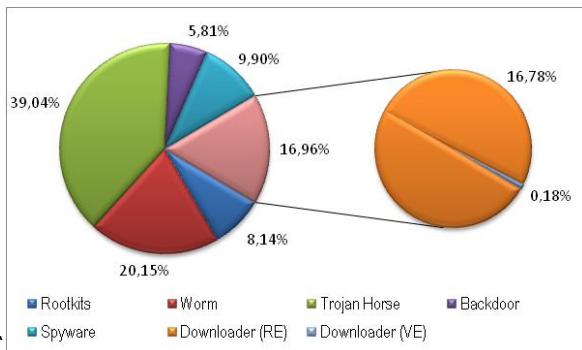


Figure 6b Percent distribution of Downloader (RE) and Downloader (VE)

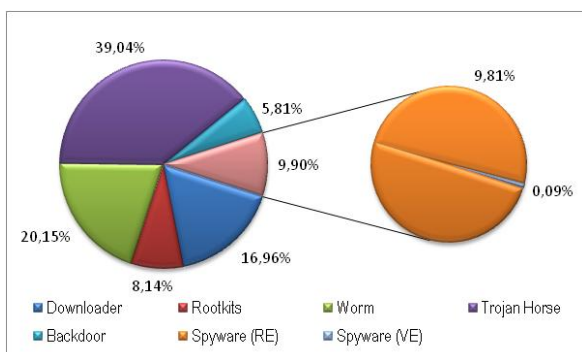


Figure 6c Percent distribution of Spyware (RE) and Spyware (VE)

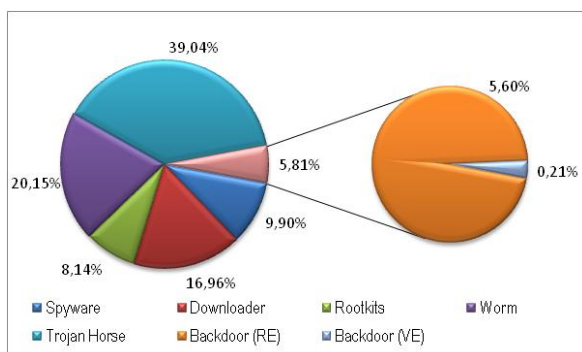


Figure 6d Percent distribution of Backdoor (RE) and Backdoor (VE)

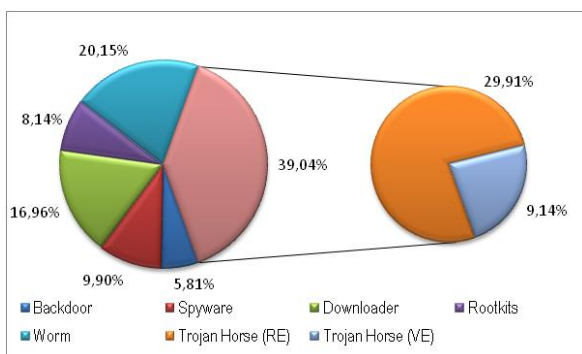


Figure 6e Percent distribution of Trojan Horse (RE) and Trojan Horse (VE)

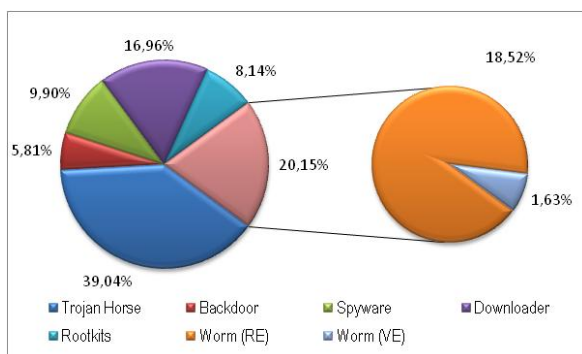


Figure 6f Percent distribution of Worm (RE) and Worm (VE)

Figure 6 Percent distribution of the attacks from the set of possible virtual attacks and their corresponding attacks from the set of successful virtual attacks

---

## Assessments and Conclusion

---

1) With respect to the virtual environment, one might say that they are a good precondition for development of new attack techniques, which can be in pawn in the malicious scenarios of the attack tools. Virtual environment can be used not only by the vendors of security software, but also by the vendors of malicious code.

2) With respect to the different techniques for attacking virtual environment, one might say that they are much and in the most cases the goal is to detect the presence of virtual environment.

3) With respect to the set of possible virtual attack, one might say:

3.1) the chosen types attack tools are appropriate for conducting analyses and making assessments;

3.2) attack tools can be used within virtual environment, but not each one has the possibility to accomplish attack to the virtual environment.

4) With respect to the set of successful virtual attack, one might say:

4.1) there are representatives from all groups of attacks and only a few are dropped down (as Computer Virus, Botnets, Dialers, and Adware);

4.2) the number of dropped down attack tools is only around 36% and one might say that the hackers are interested in using techniques for attacking virtual environment.

5) With respect to the made investigations for the percent distribution of the attack from the set of possible virtual attacks in relation to respective attacks from the set of successful virtual attacks, one might say that 17,08% of Rootkits, 1,06% of Downloaders, 0,91% of Spyware, 3,61% of Backdoors, 23,42% of Trojan Horses, and 8,09% of Worms has the possibilities to accomplish an successful attack to virtual environment.

Future analyses and investigations can be made with respect to the examination of the economical expenditure of providing security policy for different computer, system and network configurations in real environment in comparison with virtual environment.

---

## Bibliography

---

- [1] Denning, D., A Lattice Model of Secure Information Flow, Communications of the ACM, v. 19 n. 5, May 1976, pp. 236-243
- [2] Trigaux, R., A history of hacking, (<http://www.sptimes.com/Hackers/history.hacking.html>)
- [3] Shaw, W., Cybersecurity for SCADA Systems, PennWell Corp. (2006), ISBN-13: 978-1593700683, p. 194
- [4] Radhamani, G., Rao, R., Web Services Security and E-business, Global (2007), ISBN-13: 978-1599041681, p. 115, p. 25
- [5] Nickolov, E., Modern Trends in the Cyber Attacks Against the Critical Information Infrastructure, Regional Cybersecurity Forum, 7-9 October 2008, Sofia
- [6] Ferrie, P., Attacks on Virtual Machine Emulators, Symantec Advanced Threat Research, ([http://www.symantec.com/avcenter/reference/Virtual\\_Machine\\_Threats.pdf](http://www.symantec.com/avcenter/reference/Virtual_Machine_Threats.pdf))
- [7] Liston, T., Skoudis, E., On the Cutting Edge: Thwarting Virtual Machine Detection, ([http://handlers.sans.org/tliston/ThwartingVMDetection\\_Liston\\_Skoudis.pdf](http://handlers.sans.org/tliston/ThwartingVMDetection_Liston_Skoudis.pdf))

---

## Authors' Information

---

*Dimitrina Polimirova, PhD, Research Associate, National Laboratory of Computer Virology, Bulgarian Academy of Sciences, Phone: +359-2-9733398, E-mail: [polimira@nlcv.bas.bg](mailto:polimira@nlcv.bas.bg) .*

*Prof. Eugene Nickolov, DSc, PhD, Eng, National Laboratory of Computer Virology, Bulgarian Academy of Sciences, Phone: +359-2-9733398, E-mail: [eugene@nlcv.bas.bg](mailto:eugene@nlcv.bas.bg) .*

---

## DIGITAL OBJECTS – STORAGE, DELIVERY AND REUSE

Juliana Peneva, Stanislav Ivanov, Filip Andonov, Nikolay Dokev

*Abstract: The development of a methodology and tools for an automatic extraction of metadata for digital objects deployed in various subject repositories is a potential research issue. This paper presents a general overview of topics concerning the building of repositories and the applied metadata schemas with respect to the objectives of METASPEED project. The main goal of this project is to investigate and create technologies, methods and tools for automatic generation of metadata thus facilitating the proper specification of documents with different e-format, content and location.*

*Key words: Digital object repository, metadata schema*

---

### Introduction

---

Nowadays the proper supervision of organizational digital resources is very important and many companies are realizing a business advantage by managing successfully their business data. Resources are built of different kind of documents ranging from images, video or audio clips, animations, presentations, online courses, web pages, to name a few. Organizations vary in types and sizes but all of them exhibit an intensive use of digital resources because these resources are stored, distributed, shared and reused without difficulty. Certainly some barriers like technical incompatibility or missing files are to be overcome to achieve an effective use. However digital resources are increasingly being recognized as a very important organizational asset on a par with finance and human resources. So, building repositories to manage the digital content is a very important activity that brings value in the inventive deliverables of the overall organization. Each time a digital resource remains undiscovered or simply not used the organization waste time or staff efforts, misses opportunities or loses possibilities to gain a competitive advantage.

The business managerial and technical benefits of digital resources are summarized in [1]. In order to examine their value [2] and to consider the opportunities for reuse, digital resources are organized in repositories that support the organizations' policy on digital asset management. During the last five years different types of repositories ranging from digital libraries through various institutional collections and e-journals up to collaborative learning environments have been built. Each of these systems contains thousands of digital objects in the form of data and/or metadata. Content is added to a repository via different workflows and tools, and represented to the repository clients via different mechanisms. Companies as Google and Microsoft [3] are reporting for own repository investigations as well. In addition there are many workshops and the annual Open repositories [4] conference that stress on important issues concerning repository creation and management. Nevertheless the disappointments for many organizations because of the resulted greater than expected costs for set up a repository, research effort in this area appears promising.

This justifies the goal of this survey, namely to systematize some findings and discover positive research directions in this area. We are convinced that a serious step towards an increasing spread of digital object repositories comprises standardization of their content. It appears that populating a digital repository with standardized e-objects is a time and labor consuming activity. However facilitating the retrieval, sharing and using (from several users and in different context) of these objects requires their unified descriptions. Up to now

documents, disposed in electronic repositories are determined by specifications and quality metadata. Various standards depending on the subject area, e.g. SCORM, IMS, and LOM - for learning; MPEG-series for multimedia, to name a few, have been developed as well. It should be mentioned that as a rule each standard deals with a huge amount of metadata. So, the development of a methodology and tools for an automatic extraction of metadata for digital objects deployed in various subject repositories thus facilitating clients' access is a potential research issue. At the same time repositories increase successfully very quickly. Fig.1 shows the growth of the *OpenDOAR* [5] Database up to its present size and Fig.2 represents the number of repositories by country. Up to now about 1400 repositories all over the world have been reported.

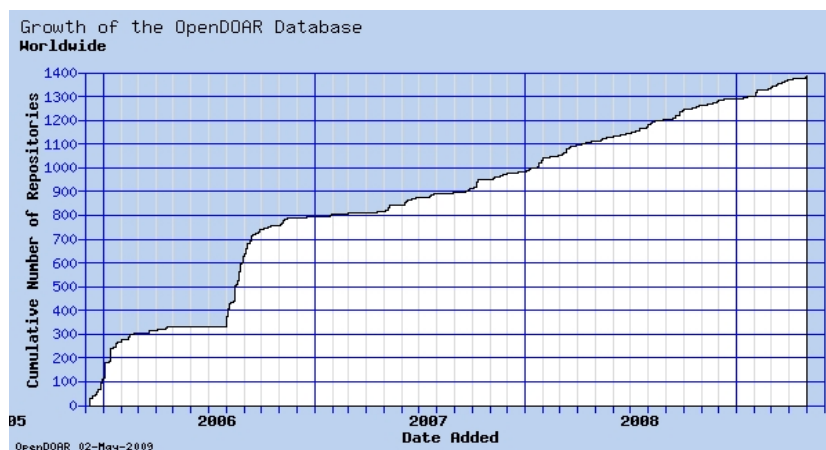


Fig. 1

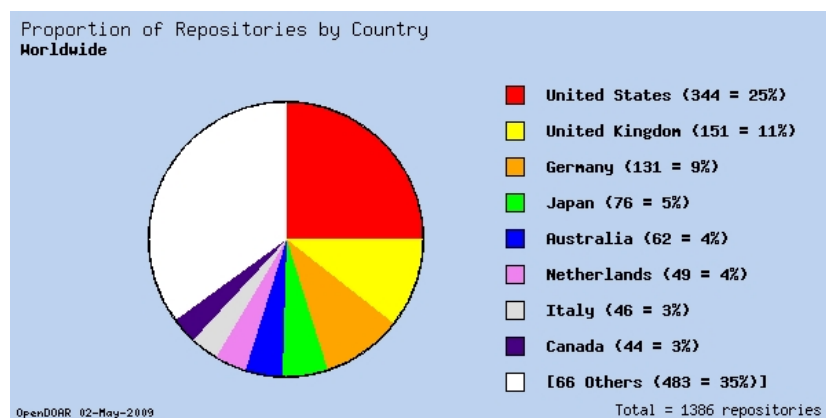


Fig. 2

In Bulgaria there are two open repositories only [6,7] registered in *OpenDOAR* 2009 and one more is forthcoming to be available at the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences [13].

Bearing these considerations in mind the paper is organized as follows: the next section introduces some of the key concepts of digital object repositories. Different classifications of repositories are briefly presented. Section 3 reviews some repository solutions. Section 4 discusses the importance of metadata and the variety of schemes/standards. Examples of widely applied schemas and their peculiarities are briefly reported. The conclusion presents the underlying project and determines some research tasks.

We have tried to do a general overview of topics concerning the building of repositories rather than to investigate particular issues in depth. Much more real work is required and it will be done within the project.

---

## Basic Definitions

---

In [8] digital objects are defined as "a data structure whose principal components are digital material, or data, plus a unique identifier for this material, called a handle (and, perhaps, other material)". This definition further evolved to capture access rules to use the object and metadata for description of the content [9]. Following these definitions digital objects can be referred as entities together with their metadata, and the services they offer to the clients.

There is no a clear definition of the concept of a repository as well. Usually any collection of digital objects is called a repository. Specialized repositories such as e-learning repositories, e-prints repositories, e-thesis repositories and subject-based repositories are being developed. However what's the difference from other datasets as directories, operational databases, catalogues, and portals? Defining institutional repositories as a "set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members" Lynch [10] focuses more on the services that a repository is supposed to offer. The type of the content and the different technological solutions e.g. free versus commercial software remain in the background or even lose to. The organizations differ in the underlying motivation to build digital repositories as well. Services that are expected from repositories range across several functional areas depending on the interest for different communities (digital libraries, research, learning, e-science, publishing, records management, preservation). Among these functional areas data sharing, preservation of digital resources, corporate information management and scientific collaboration are to be considered. Not surprisingly there is no a unique definition of the notion "digital object repository". During a special meeting of leading companies and associations [11] a repository has been defined as "a networked system that provides services pertaining to a collection of digital objects". Following this general definition repositories include: institutional collections, datasets, learning objects banks, cultural heritage artifacts, etc. Generally speaking a digital repository can be considered as means of handling digital content. Thus they may include a wide range of content for a variety of purposes and users. What goes into a repository depends on decisions made by each institution or administrator. The peculiarities of digital repositories that distinguish them from other digital collections are summarized in [12]. In addition an attempt to develop a classification of repositories is also proposed. According to Heery and Anderson [12] repositories can be typified by content (corporate records, e-theses, learning objects, research data), by coverage (personal, institutional, national, journal), by users (learners, researchers, teachers, etc.) and by function (access, preservation, dissemination, reuse). In [14] two more features of the repositories, namely policy (persistence, deposit, access) and infrastructure – centralized versus distributed have been taken into account. It is very important to determine the content and scope of any repository because this is the way to define the managerial policies. *OpenDOAR* provides a list of major content types (publications, books, conference papers, theses, learning objects, multimedia items, etc.) as well. Some definitions of the various types are given in [15].

In our opinion the content type of the repositories (cultural heritage, e-learning objects and teaching materials, scientific papers, e-maps, etc.) makes a good distinction about the main groups of users and corresponding managerial policies.

---

## Repository Solutions

---

Digital repository solutions consist of hardware, software and open standards. Recently the more commonly adopted software solutions fall into two broad groups: open source and commercial software.

Open source software is exemplified by DSpace, Fedora, and EPrints. DSpace is the software of choice for academic, non-profit, and commercial organizations building open digital repositories. DSpace preserves and enables easy and open access to all types of digital content including text, images, moving images, mpegs and data sets. It is applied for accessing, managing and preserving scholarly works [16]. Fedora (Flexible Extensible Digital Object Repository Architecture) was originally developed by researchers at Cornell University as an architecture for storing, managing, and accessing digital content in the form of digital objects [8]. Nowadays the Fedora Repository Project and the Fedora Commons community together with the DSpace project are under the supervision of the not-for-profit organization DuraSpace [17]. The Fedora Repository Project (simply Fedora) implements the Fedora abstractions [18] and provides basic repository services. This permits to express digital objects, to assert relationships among digital objects, and to link services to digital objects. Fedora ensures the durability of the digital content by providing tools for digital preservation. The Fedora Commons community deals with producing additional tools and applications that enlarge the functionality of the Fedora repository. The latter is extremely flexible and can be used to support any type of digital content. There are numerous examples of Fedora being used for digital collections, e-research, digital libraries, archives, digital preservation, institutional repositories, open access publishing, document management, digital asset management, and more [18]. Fedora Commons provides sustainable technologies to create, manage, publish, share and preserve digital content. EPrints [19] is an open source platform for building repositories of research literature, scientific data, and student theses. There are now over 210 repositories using the EPrints software, the repository at New Bulgarian University being one of them.

Commercial software could be based on an open source repository engine coupled with a proprietary application software layer – VITAL [20]. Other possibility includes openly accessible API's using XML interfaces - DigiTool [21] and DPS [22]. Because of the increased demand to manage digital assets, libraries need standard methods and tools to facilitate cataloging, sharing, searching, and retrieval of digital collections. Through highly customizable user interfaces DigiTool enables academic libraries and library consortia to manage and provide access to the growing volume of digital collections. Support for library standards and built-in integration with other Ex Libris products, e.g., Aleph®, Voyager®, MetaLib®, SFX®, and Primo, makes DigiTool an integral part of the library infrastructure and facilitates the incorporation of digital resources into library services. VITAL is an institutional repository solution built on Fedora, It is designed to simplify the development of digital object repositories and to provide online search and retrieval of information for administrative staff, contributing faculty and end-users. VITAL provides all functions such as storing, indexing, cataloging, searching and retrieving required for handling large text and rich content collections.

A functional comparison of repository software products is presented in [23]. Consulting services are available through Sun [24].



---

## The Importance of Metadata

---

In order to be easily retrieved, shared and used from different users and for different purposes, various types of e-documents have to be described following common schemas and rules e.g. specifications/standards and metadata. The term metadata e.g. data about data is used differently ranging from machine understandable information through records that describe electronic resources. In a library, "metadata" applies for any kind of resource description. Metadata describe how and when and by whom a particular set of data was collected, and how the data is formatted. Metadata is essential for understanding information stored in data warehouses and has become increasingly important in XML-based Web applications [27]. In addition they ensure the accessibility, identification and retrieval of resources. Descriptive metadata facilitate the resources' organization, interoperability and integration, provide digital identification and support archiving. Poor quality or non-existent metadata mean that resources remain invisible within a repository or archive thus becoming undiscovered and inaccessible. In the case of digital assets, metadata usually are structured textual information that describes something about the creation, content, or context of an image [28].

There are several types of metadata:

1. Descriptive - title, author, extent, subject, keywords
2. Structural – unique identifiers, page numbers, special features (table of contents, indexes)
3. Technical - file formats, scanning dates, file compression format, image resolution
4. Preservation - archival information
5. Rights management - ownership, copyright, license information

Metadata can be stored in different ways:

1. Separately as a HTML, XML or MARC21 (format for library catalogues) document linked to the resource
2. In a database linked to the resource
3. As an integral part of the record in a database or embedding the metadata in the Web pages

Nevertheless the importance of metadata has been recognized, means for efficient implementation still lack. Due to the rapid growth in digital object repositories and the development of many different metadata standards metadata implementation is complex. On the other hand quality metadata can be produced by experts in the subject domain only. So far, most of the resource discovery metadata are still created and corrected manually either by authors, depositors and/or repository administrators. It appears attractive to auto-generate metadata with no human intervention. Recent research findings are reported in [25, 26].

In order metadata to be processed via computer, proper encoding has to be applied. This is done by the addition of markup to a document to store and transmit information about its structure, content or appearance. Schemas comprise metadata elements designed to describe particular information. We can mention the following encoding schemas concerning how metadata is presented:

1. HTML (Hyper-Text Markup Language)
2. XML (eXtensible Markup Language)
3. RDF (Resource Description Framework)
4. MARC (Machine Readable Cataloguing)
5. SGML (Standard Generalized Markup Language)

Metadata schemas can be viewed as standards describing the categories of information to be recorded. They ensure consistency in metadata application, support interoperability of applications and resource sharing. Schemas are built from individual components, i.e. metadata elements. Depending on the element definition each element contains a particular category of information. Certainly not all schemas contain the same elements as the needs of users differ.

There are widespread metadata standards (schemes) that are used in digital object repositories [29, 30]. Standards are being developed all the time. Below some well-known examples are listed.

1. Dublin Core [31]. The Dublin Core standard arose from a 1995 workshop held in Dublin, Ohio. The basic DCMES (Dublin Core Metadata Element Set) involves 15 elements. Each is optional and repeatable, and may appear in any order the creator of the metadata wishes. This simple generic element set is applicable to a variety of digital object types. It is used for the description of simple textual or image resources. For richer descriptions to enable more refined resource discovery, Qualified Dublin Core has been developed. This standard employs additional qualifiers to the basic 15 elements to further refine the meaning of an element. Qualifiers increase the precision of the metadata.
2. TEI (Text Encoding Initiative) [32]. The Text Encoding Initiative is an international project to develop guidelines for marking up electronic texts such as novels, plays, and poetry, primarily to support research in the humanities.
3. METS (Metadata Encoding & Transmission Standard) [33]. METS is maintained by the Network Development and MARC Standards Office of the Library of Congress. The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library expressed using the XML schema language.
4. MODS (Metadata Object Description Schema) [34]. This is an XML schema for descriptive metadata compatible with the MARC 21 bibliographic format. It includes a subset of MARC fields and uses language based tags rather than the numeric ones used in MARC 21 records. In some cases, it regroups elements from the MARC 21 bibliographic format. Like METS, MODS is expressed using the XML schema language.
5. EAD (Encoded Archival Description) [35]. (EAD) was developed as a way of marking up the data contained in finding aids so that they can be searched and displayed online. In archives and special collections, resources are described via a finding aid. Finding aids differ from catalog records by being much longer, more narrative and explanatory, and highly structured in a hierarchical fashion. They generally start with a description of the collection as a whole, indicating what types of materials it contains and why they are important. The finding aid describes the series into which the collection is organized and ends with an itemization of the contents of the physical boxes and folders comprising the collection.
6. LOM (Learning Object Metadata) standard (IEEE 1484.12.1-2002) [36]. LOM was developed by IEEE Learning Technology Standards Committee to enable the use and re-use of technology-supported learning resources such as computer-based training and distance learning. Learning Objects are defined here as any entity, digital or non-digital, which can be used, re-used or referenced during technology supported learning. The LOM defines the minimal set of attributes to manage, locate, and evaluate

---

learning objects. Where applicable, Learning Object Metadata may also include pedagogical attributes such as; teaching or interaction style, grade level, mastery level, and prerequisites.

7. MARC Standards [37]. The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form. Today this is the most common standard format used by bibliographic and library catalogues to exchange information on their contents.
8. VRA Core Categories [38]. This is a scheme developed by the Visual Resources Association for the description of art, architecture, artifacts, and other visual resources. It consists of a metadata element set (units of information such as title, location, date, etc.), as well as an initial blueprint for how those elements can be hierarchically structured. The element set provides a categorical organization for the description of works of visual culture as well as the images that document them.
9. MPEG standards [39]. The ISO/IEC Moving Picture Experts Group (MPEG) has developed a suite of standards for coded representation of digital audio and video. Two of the standards address metadata: MPEG-7, Multimedia Content Description Interface (ISO/IEC 15938), and MPEG-21, Multimedia Framework (ISO/IEC 21000). MPEG-7 defines the metadata elements, structure, and relationships that are used to describe audiovisual objects including still pictures, graphics, 3D models, music, audio, speech, video, or multimedia collections. MPEG-21 was developed to address the need for an overarching framework to ensure interoperability of digital multimedia objects.
10. CSDGM (Content Standard for Digital Geospatial Metadata) [40]. This is a metadata schema for geospatial datasets comprising topographic and demographic data, GIS (geographic information systems), and computer-aided cartography base files. Geospatial datasets are used in many areas e.g. land use studies, biodiversity counts, climatology and global change tracking, remote sensing, and satellite imagery. An international standard, ISO 19115, Geographic Information— metadata was issued in 2003.

---

## Conclusion

---

In this paper we examine some issues concerning digital object repositories and metadata. Besides the possibility of applying some well-known world standards such as SCORM, IMS, MPEG-7, it is not expected that the shared e-documents will be specified in a uniform way. This justifies any research effort to raise the productivity of the process of metadata generation for different e-documents. Taking into account the rapidly growing number of new digital repositories investigations in this area are promising.

Metadata ExTraction for Automatic SPEcifications of E-Documents – METASPEED is a Bulgarian research project funded by the Bulgarian National Science Fund under the thematic priority: Information and Communication Technologies. It aims to facilitate the development of Bulgarian standards and even commonly accepted specifications for the description of metadata for e-documents in different subject areas.

The goal of this project can be briefly summarized as follows: to investigate and create technologies, methods and tools for automatic generation of metadata thus facilitating the proper specification of documents with different e-format, content and location.

Project findings will facilitate the access to different digital collections in a straightforward manner. This is the first stage toward the development of an integrated information environment in Bulgaria. We expect that the main contributions will include:

- 
- development of proper tools for an automatic metadata generation for collections containing digital documents of different shapes and types;
  - building a framework to share European and Bulgarian e-resources;
  - development of national standards for document sharing.

The work underlying the METASPEED project comprises:

- survey of standards and schemas for metadata;
- evaluation of current automatic metadata extraction and generation tools;
- compilation of recommended functionalities for automatic metadata generation applications;
- investigations on specialized technologies and methods for metadata retrieval for documents in different areas;
- design and implementation of proper software prototypes;
- prescriptions for Bulgarian standards in different subject areas;
- a methodology how to build an integrated information repository for digital documents in Bulgaria.

---

## Acknowledgements

This work is partially granted by Bulgarian National Science Fund, Ministry of Education and Sciences in the frame of the project "Automated Metadata Extraction for e-documents Specifications and Standards", contract No: D002(TK)-308/ 19.12.2008.

---

## Bibliography

- [1] Duncan C. Digital Object Repositories Explained, an Intrallect White Paper, 2006
- [2] Bluth E., Chandra V. The Value Proposition in Institutional Repositories EDUCASE Review, September/ October 2005.
- [3] <http://dorsdl2.cvt.dk/>
- [4] <http://openrepositories.org/>
- [5] <http://www.opendoar.org/> - Directory of Open Access Repositories
- [6] <http://eprints.nbu.bg/> – New Bulgarian University Scholar Electronic Repository
- [7] <http://research.it.fmi.uni-sofia.bg:8880/dspace/> - Research at Sofia University
- [8] Kahn R., Wilensky R. A Framework for Distributed Digital Object Services, 1995, <http://www.cnri.reston.va.us/home/cstr/arch/k-w.html>
- [9] Lagoze Carl. 1995. A Secure Repository Design for Digital Libraries. D-Lib Magazine, <http://www.dlib.org/dlib/december95/12lagoze.html>
- [10] Lynch, C. (2003). "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age." ARL, 226, February 2003, 1-7, <http://www.arl.org/resources/pubs/br/br226/br226ir.shtml>
- [11] <http://msc.mellon.org/Meetings/Interop/terminology.doc>
- [12] Heery R., Anderson, S.: Digital Repositories Review, AHDS, 2005.
- [13] <http://www.driver-support.eu/pmwiki/index.php?n=Main.Bulgaria>
- [14] <http://www.ukoln.ac.uk/repositories/digirep/index/Typology>
- [15] <http://www.rsp.ac.uk/content>

- 
- [16] <http://www.dspace.org/>
- [17] <http://duraspace.org/>
- [18] <http://www.fedora-commons.org/>
- [19] <http://www.eprints.org/>
- [20] <http://www.vtls.com/products/vital>
- [21] [www.exlibrisgroup.com/digitool.htm](http://www.exlibrisgroup.com/digitool.htm)
- [22] [www.exlibrisgroup.com/Preservation.htm](http://www.exlibrisgroup.com/Preservation.htm)
- [23] <http://www.rsp.ac.uk/repos/software/surveyresults>
- [24] Grant C. Delivering digital repositories with open solutions, a Sun white paper, Version 8.0, November 2007.
- [25] Polfreman M. and Rajbhandaji S. Metatools – Investigating Metadata Generation Tools, JISC Final report, Oct.2008.
- [26] Greenberg J. et al. Final Report of the AMEGA Project, UNC School of Information and Library Science, 2005.
- [27] <http://www.webopedia.com/TERM/M/metadata.html>
- [28] <http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/metadata-overview/>
- [29] <http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/putting-things-in-order-links-to-metadata-schemas-and-related-standards/>, February 2009
- [30] <http://metadata-standards.org/>
- [31] <http://dublincore.org/>
- [32] <http://www.tei-c.org/>
- [33] <http://www.loc.gov/standards/mets/>
- [34] <http://www.loc.gov/standards/mods/>
- [35] <http://www.loc.gov/ead>
- [36] <http://ltsc.ieee.org/wg12/>
- [37] <http://www.loc.gov/marc/marc.html>
- [38] <http://www.vraweb.org/projects/vracore4/index.html>
- [39] <http://www.mpeg.org/>
- [40] [http://www.fgdc.gov/metadata/index\\_html](http://www.fgdc.gov/metadata/index_html)
- 

### Authors' Information

---

*Juliana Peneva* – New Bulgarian University, Department of Informatics; e-mail: [july\\_peneva@abv.bg](mailto:july_peneva@abv.bg)

*Stanislav Ivanov* – New Bulgarian University, Department of Informatics; e-mail: [sivanov@nbu.bg](mailto:sivanov@nbu.bg)

*Filip Andonov* – New Bulgarian University, Department of Informatics; e-mail: [fandonov@nbu.bg](mailto:fandonov@nbu.bg)

*Nikolay Dokev* – New Bulgarian University, Department of Informatics; e-mail: [n.dokev@nbu.bg](mailto:n.dokev@nbu.bg)

---

## MULTI-MODAL EMOTION RECOGNITION – MORE "COGNITIVE" MACHINES

Velina Slavova, Hichem Sahli, Werner Verhelst

*Abstract:* Based on several results related to studies on emotions, we suggest that the process of emotion-recognition is assisted by some internal structure of the cognitive images of emotions, which are at different levels of knowledge representation. We concede that the main proposed in psychology models are in correspondence with these levels and in this sense - complementary. We summarize the state-of-the-art of machine emotion recognition with regards of the used psychological models of emotions. In order to discover amelioration sources of multimodal machine emotion recognition, we propose a scheme of the cognitive process, based on gradual levels of representation. The proposed scheme shows several "strategic" differences with the architectures used in machine emotion recognition. We discuss the questions related to recognition, assisted by two levels of representation that we called "perceptual" and "conceptual".

*ACM Classification Keywords:* I.2 Artificial Intelligence, 1.2.0.Cognitive simulation

---

### Introduction and Development of Previous Hypothesis

---

Recent advances in human-computer interaction (HCI) show the importance of applying knowledge from different disciplines in order to make machines more "intelligent". The wish to produce machines with more and more "intelligence" necessitates providing them the capacity of sensing human emotions. During the last years the research concentrated in all these problems. An example of that is the Human-Machine Interaction Network on Emotion (HUMAINE) – a network of excellence that aims to lay the foundations for European development of systems that can register, model and/or influence human emotional and emotion-related states and processes - 'emotion-oriented systems'. [<http://emotion-research.net/>].

*Emotion* is a mental<sup>1</sup> and physiological state associated with a wide variety of feelings, thoughts, and behaviour. The associated physiological state is often accompanied by physiological changes which can lead to modifications in persons' observable and measurable manifestations, called expressions. Expressions can be perceived and evaluated by others as evidences of given emotional state.

Emotions are subject of study in several disciplines - psychology, cognitive science, philosophy and computer science. Although there have been numerous studies with regards to both the psychological and the computational aspect of emotions, it is still not clear how to define and how to categorize them. There are two basic theoretical "models" coming from psychology. The first model is "discrete" (Fig. 1, A). Emotion categories are determined as entities with names and descriptions. Several researchers (see Ekman 1992) argue that a few emotions are basic or primary (*anger, disgust, fear, joy, sadness, and surprise*). This approach is very convenient for implementation purposes as it provides "ready-on" classes for the machine learning and classifiers. The second model represents emotions as having certain *properties* in a continuous space, on axes as pleasant–unpleasant, active–passive, attention–rejection, simple–complicated, etc. (ex. Schlosberg 1954). Two commonly used dimension-axes are "*valence*" and "*arousal*" (fig. 1, B), where valence describes the pleasantness of the stimuli. Although the consensus of the experts from HUMAINE is that "labelling schemes based on traditional

---

<sup>1</sup> The term *Mental state* is often defined as "state of a person's cognitive processes".



As shown in figure 3, subjects' evaluation shapes clusters that cover more general areas of the plane. We raise the hypothesis about the existence, on the perceptual level of representation, of more general emotional categories, as shown in figure 4. The claim is that the conceptual knowledge about emotions is structured on two levels of representation ("perceptual" and "conceptual") and that its internal structure is highly explored in emotion recognition tasks.

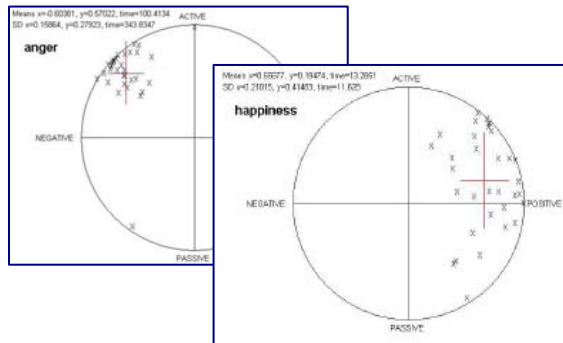


Fig 3. Evaluation of "anger" and "happiness" [Wan et al., 2005]

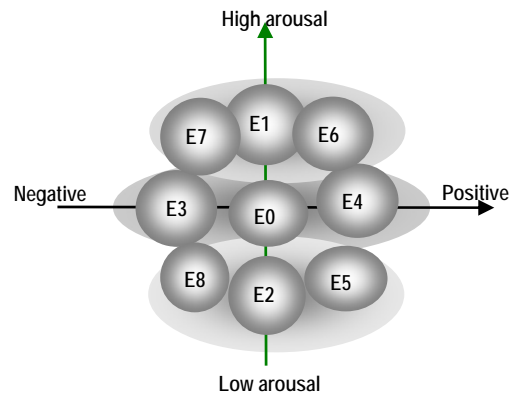


Fig 4. General categories on the "perceptual level"

In the next section we give a short overview of machine emotion recognition with regards of the used models. Section 3 proposes a processing architecture which relates the two levels and the last concludes the paper.

### A Trial to Over-fly Machine Emotion Recognition

Extensive surveys in areas such as facial expression analysis (Fasel and Luetttin 2003) vocal emotion (Oudeyer 2003), gesture recognition (Turk, 2001; Marcel 2002, Heylen (2006)) head tracking (Jaimes and Sebe 2005) have been published during the last decade. The problem of multimodal emotion recognition approaches are extensively discussed in very recent surveys (Sebe et al. 2005, Jaimes and Sebe 2007, Zeng et al. 2009). Our attempt is to make a parallel between the existing methods of machine emotion classification and a generalized scheme of the process of multimodal emotion recognition by humans.

**Facial Expression Recognition.** Studies, that are known in the domain (Ekman and Friesen 1978, Ekman 1989, Ekman 1992), have emphasized that facial expressions are universally expressed and recognized by humans. They focused on a set of seven emotions that have associated facial expressions (*fear, anger, sadness, happiness, disgust, surprise, and contempt*). The so-called Facial Action Coding System was developed to code facial expressions where movements on the face were described by a set of Action Units. Ekman's work inspired many researchers to use image and video processing in order to classify emotions. Some methods follow feature-based approach (tracking of specific features such as the corners of the mouth, eyebrows, etc.) and other use a region-based approach (facial motions are measured in certain regions such as the eye/eyebrow and the mouth). Two types of classification schemes are used: dynamic and static. Dynamic classifiers (Hidden Markov Models HMM) use several video frames and perform classification by analyzing the temporal patterns of the regions analyzed or features extracted. Static classifiers classify each frame in a video to one of the facial expression categories based on the results of a particular video frame. These methods are similar in the general sense that they first extract some features from the images, then these features are fed into a classification system, and the outcome is one of the pre-selected emotional categories.



Following the analyses by Sebe (2005), the performance of the existing systems varies between 74% and 98% depending on the algorithm, the number of emotional categories etc. More details are given in the survey by Jaimes and Sebe (2007), where the authors summarize that the used approaches suffer from the following limitations: 1) they handle a small set of posed prototypical facial expressions of six basic emotions from portraits or nearly frontal views of faces; 2) they do not perform a context-dependent interpretation of shown facial behavior; 3) they do not analyze extracted facial information on different time scales and can't analyze mood and attitude (larger time scales). On-going works are trying to overcome some of these limitations. For example, Hu et al. 2008 generated multi-view images of facial expressions from 3D data and showed that non-frontal views give better results in the developed system for emotion recognition.

It has to be pointed out that the majority of the developed systems use a set of basic emotions. Interestingly, practical results for person independent emotion recognition have shown that the classification rates are significantly higher when considering more general categories such as "*positive*", "*negative*", "*neutral*" and "*surprise*" (Sebe et al. 2002). Recently, some authors have made some attempts to generalise the obtained results to more global categories. E.g, Zeng et al. 2006 have developed a classifier for realistic data which categorizes directly to positive, negative and non-emotional states, simply using labeled training data to these three general categories. That corresponds very much to the general perceptual categories, proposed in figure 4. To our knowledge, there are not theoretical works in the direction of relating facial expressions with emotion dimensions (properties) or with generalized categories.

*Speech emotion recognition.* The systems for speech emotion recognition use techniques for extraction of relevant characteristics from the raw speech signal. Acoustic correlates of basic emotional categories are investigated in terms of pitch, energy, temporal and spectral parameters, on different time-scales, etc... with the aim to extract emotion-relevant information (for recent advances - consult the site of EU-IST Network of Excellence HUMAINE <http://emotion-research.net/>). The data-driven approaches for recognition use supervised machine learning algorithms (neural networks, support vector machines, HMM etc.) that are trained on databases of emotional speech, containing labelled utterances with a previously chosen set of basic emotions. One main problem of speech emotion recognition is related to the "training corpus" dependency of the classifiers, as discussed in Slavova, Verhelst and Sahli 2008. Other problems are related to 1) the bad performance of the classifiers in noisy environments and to 2) their speaker dependency.

In general, the reported results (obtained in laboratory conditions, within the corpora used for training) the recognition accuracy of the machine classifiers is comparable with the human "acoustic" categorization capacities (around 66% for 6 basic emotions following Scherer et al. 2001). Several systems use in addition language semantics to improve the results. We provide a few examples: Muler et al. (2004) report that the fuse of acoustic and language information increased the recognition rates from 74% to 92%. Chuang and Wu (2004) applied a keyword spotting system in order to transform the speech signal into textual data and report that the used 500 keywords played a decisive role in the "outside" test, where the acoustic module could not perform satisfyingly. All these strategies require speech recognition and charge the systems with additional complexity.

To our knowledge, the existing systems for speech emotion recognition are based on the discrete model. However, there are results at finding a correspondence between the audio-signal and the emotion-dimensions model. Research in speech synthesis has shown that nearly all acoustic variables show substantial correlations with the emotion dimensions (see for example Schröder M. et al. 2001, Schröder M. 2004).

---

*Recognition using biological signals.* Contemporary studies showed that parameters from measurements as electrocardiograms (ECG), electromyography (EMG), electroencephalograms (EEG), respiration and skin conductivity, are highly correlated with the emotional states. The first trial (Fridlund and Izard 1983) to apply pattern recognition to classification of 4 basic emotions using EMG features attained rates of 38-51% accuracy. Since then, the results are considerably better. Takahashi (2004) applied support vector machines in a classifier using data from EEG, pulse, and skin conductance and obtained very promising recognition rates for three of 5 emotions (around 70% for *joy*, *anger*, and *fear*). Nasos et al. (2004) used wireless sensors for measuring temperature, heart rate and galvanic skin response and obtained promising results for *fear*, *sadness* and *anger* (and not so good for *amusement*, *surprise* and *frustration*). Lisetti and Nasos (2004) used galvanic skin response, heart rate and temperature and showed that three different supervised learning algorithms can generalize from new collections of signals. Despite of the evidences (Kim 2004), showing that biological changes are highly correlated with the emotion dimensions, these systems are conceived on the bases of the discrete model, i.e., the first step is to construct data models (EEG, ECG, temperature etc.) that correspond to basic emotions. Yet, the use of emotion dimensions is possible and gives promising results. Nakasone et al. (2005) developed a system based on the valence-arousal model which realizes real-time recognition of emotions in a game between human user and humanoid agent, using data from EMG and skin conductance. The system, based on Bayesian network, discriminated well different *arousal levels*, indicated by Galvanic skin response, but had difficulties with the *negatively valenced* emotions, indicated by EMG.

**Multimodal emotion recognition.** The first attempt (Huang et al. 1998) to use an "audio-visual" feature vector increased the performance and some confusions, made by a single-modality classifier were resolved. The results, obtained by all other realized audio-visual emotion recognition systems are in the same direction. We may cite the works of Chen (2000), Yoshitomi et al. (2000), De Silva and Ng (2000), Go et al. (2003), Schuller et al. (2004), Zeng et al. (2004), Song et al. (2004), Sebe et al. (2006), and Zeng et al. (2007) who investigated the effects of a combined detection of facial and vocal expressions of affective states using different types of techniques and classifiers (HMM, SVM, Bayesian networks etc). In brief, these systems achieve an accuracy of 72–85% when recognising basic emotions from clean audiovisual input (e.g., noise-free recordings, closely placed microphone, non-occluded portraits) from an actor speaking a single word and showing exaggerated facial displays of a basic emotion. In their extended survey, Jaimes and Sebe (2007) note that the developed audio-visual systems *have most of the drawbacks of the unimodal analyzers*.

An number of efforts are reported toward multimodal systems based on other data sources, for example using haptical interaction on a touch-screen, via the mouse etc. (see Schuller et al. (2002) Maat and Pantic (2006)).

**Multimodal Problems.** Important problems are related to data fusion. First, "architectural" problems, related to the stage on which the fusion should be realised. The survey of Sebe et al. (2005) reports that multi-sensory data are typically processed separately and only combined at the end. In fact, several works recommend early fusion, for example Chen (2000) argues that to realize human-like multimodal analysis, the input data should be processed in a joint feature space. Second - "processing problems" - the size of joint feature space, the different feature formats and timing. These problems are discussed in details in Sebe et al. (2005).

Other problems are related to the multimodal training data. They concern the labelling, the use of unlabeled data, spontaneous or posed collection etc. The survey by Zeng et al. (2009) points that the existing methods typically handle deliberately displayed, exaggerated expressions of prototypical emotions despite the fact that deliberate behaviour differs in visual appearance, audio profile, and timing from spontaneously occurring behaviour.

In conclusion, effective emotion recognition is likely to take place when different input devices <sup>1</sup> are used in combination. Nevertheless, the multimodal machine emotion recognition is still in its infancy. It is clear that human subjects perform the task much better, in real time, in real situation, in noisy environment, looking from different angles. They use additional sources of information such as context and semantics of the produced speech, as well as "knowledge" about their own mind when estimating other peoples' states (a phenomenon, known as mind-reading). Our presentiment is that the improvement of recognition systems necessitates more fundamental analyses in the direction of the used theoretical models of emotions, with regards of the inter-related levels of representation (fig. 1), phases of recognition, interactions between the modality-specific information flows etc.

**Multimodal Architecture – Some General Comments**

In the proposed approach emotion recognition by human subjects is considered as a process of mapping of obtained from external sources information to existing conceptual knowledge. The process concerns treatment by sensory systems. As known, a *sensory system* is a part of nervous system consisting of sensory receptors that receive stimuli from environment, neural pathways that conduct this information to brain and parts of brain that processes this information. For the modeling purposes, we propose a general scheme of this processing, given on fig. 5.

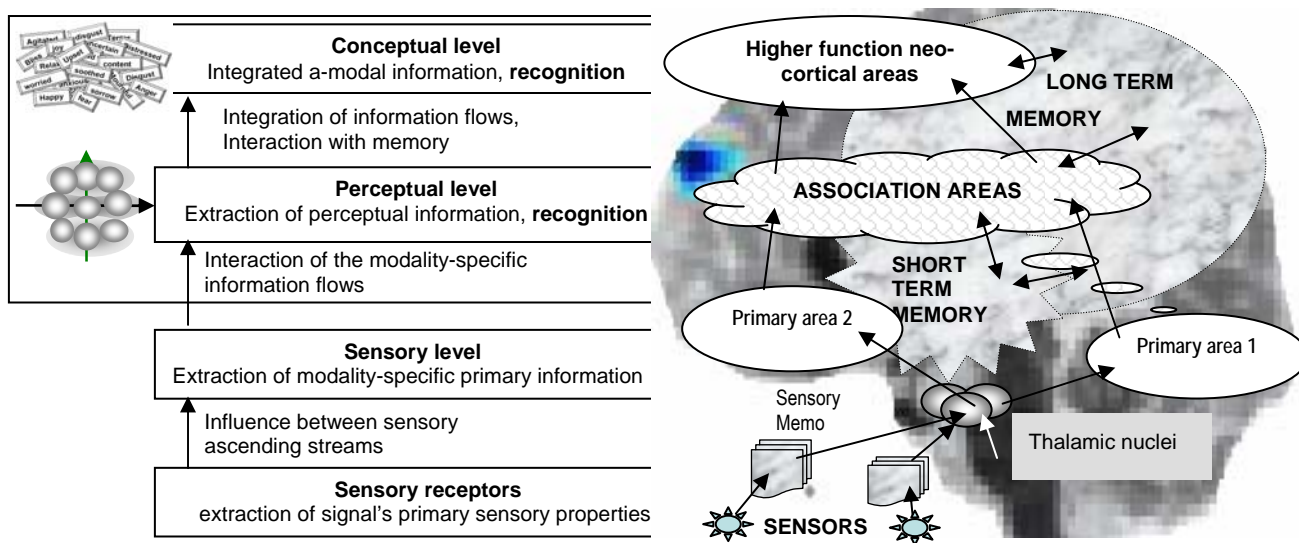


Fig. 5. General scheme of multi-modal processing

The sensors extract some primary features from the stimuli and keep short-time paths of the flow in the sensory memory. The information is conducted to different thalamic nuclei where the sensory flows influence each other. The flows are separately conveyed to the primary (modality specific) neo-cortical areas, which perform to obtain a novel, modality-specific representation. The resulting information flows are transmitted to separate association

<sup>1</sup> Several efforts are directed to machine emotion recognition based on **gestures**, using data from head and body movements, gaze etc. We have not included this direction in the proposed over-flight of the entire domain, we give only some references

(secondary) neo-cortical areas, where they interact to obtain novel, separate representations. On this stage are recognized meaningful perceptual features, in interaction with memory. That corresponds on the "perceptual level", given on fig.1. On the final stage, an a-modal conceptual representation is obtained, in result of integrating relevant information from of all the ascending flows and in interaction with memory.

Once the obtained, by bottom-up processing, information is identified to existing knowledge (on perceptual and conceptual level), a top-down process starts, which, briefly, propagates the features of the identified concepts on all the levels of the sensory hierarchy and the system "concentrates" on "expected" features.

None of the systems for machine emotion recognition is based on principles, similar to the above described scheme. Here we concentrate on the fact that human memory stores at time models of the emotion categories and of the emotion dimensions (properties). Our hypothesis is that two stages of recognition coexist, – at the "perceptual" and at the "conceptual" levels, which is important for the recognition strategy. On the left part of figure 5 we propose a general scheme of the processing. In summary, the information which can be compared to memory models is obtained at several steps. That includes 1. uni-modal feature extraction with memory-storage, 2. classification to general perceptual categories. 3. evaluation of the perceptual image using appropriate weights for the features of the input modalities, 4. identification of the perceptual category, 6. generation of a resulting feature-vector, 7. categorization on conceptual level using context, 8. Inducing of top-down process. Each step necessitates further clarification as well as modeling of interactions with memory.

---

## Conclusion, On-going and Future Work

---

In conclusion, none of the existing systems for machine recognition uses multiple successive stages of fusing information in novel informative representations in order to recognize different levels of knowledge. We suppose that the recognition on the "perceptual" level is an important step which can not be omitted. Moreover, it seems that real-time, person independent, realistic data etc machine recognition, is more reliable when more global categories are used, as in the proposed here structure of the "perceptual level".

With the intention to realize multi-modal emotion recognition, we assume that a lot of efforts have to be performed for conceiving a system, working satisfyingly in realistic environment. When taking into account the provided the presented approach, we concenter that: The uni-modal recognition should be reviewed with concerns of the general categories of the "perceptual" level (the number of dimensions is also to be clarified). Several studies show that the information from audio and from video has different "weights" of recognition importance depending on the emotion displayed. On-going research tries to clarify these dependencies. Our belief is that, when combining the results and with an appropriate use of computational techniques, such general strategy will give more robustness of the system and will lead to a reduction of used feature-vectors. A special effort has to be done in discovering the relations between the two levels of representation. That includes also an appropriate modeling of the top-down influence in attempt to concentrate the computational resources on relevant information. Of course, all these questions cannot be solved without an appropriate model of memory and necessitates being tested and adjusted in implemented machine realizations.

---

## Acknowledgements

---

The paper is partially financed by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA and the Consortium FOI Bulgaria. [www.ithea.org](http://www.ithea.org), [www.foibg.com](http://www.foibg.com)

---

**Bibliography**

---

- Barsalou L. W., Simmons W. K., Barbey A. K. and Wilson C. D., 2003, Grounding conceptual knowledge in modality-specific systems, in: Trends in Cognitive Sciences, Volume 7, Issue 2, 2003
- Chen L. and Huang T. 2000, Emotional expressions in audiovisual human computer interaction, in Proc. International Conference on Multimedia and Expo (ICME), 2000.
- Chen L.S. 2000, Joint Processing of Audio–visual Information for the Recognition of Emotional Expressions in Human–computer Interaction, PhD thesis, Univ. of Illinois at Urbana-Champaign (2000).
- Chuang Ze-Jing and Chung-Hsien Wu, 2004, Multi-Modal Emotion Recognition from Speech and Text, in: Computational Linguistics and Chinese Language Processing, Vol. 9, No. 2 , 2004
- De Silva, L.C. and Ng P.C., 2000, Bimodal emotion recognition, in Proc. Face and Gesture Recognition Conf., 2000.
- Ekman Paul and Friesen W. 1978, Facial Action Coding System: Investigator's Guide, Consulting Psychologists Press, 1978.
- Ekman, Paul 1992. Are there Basic Emotions? Psychological Review, 99(3): 550-553.
- Fasel B.,Luettin J. , 2003, Automatic facial expression analysis: a survey, Pattern Recognition 36 (2003)
- Fridlund A. and Izard C., 1983, Electromyographic studies of facial expressions of emotions and patterns of emotion, in : Social Psychophysiology: A Sourcebook, J. Cacioppo and R. Petty, eds., 1983.
- Heylen, D.K. (2006), Head gestures, gaze and the principle of conversational structure, International Journal of Humanoid Robotics, Vol: 3 Issue: 3, 2006
- Hu Y, Zhihong Zeng, Yin L., Wei X., Tu J. and Huang T.S., 2008, A Study of Non-frontal-view Facial Expressions Recognition, in: proc of the 19th International conference on pattern recognition, 2008.
- Huang TS, Chen LS, Tao H, Miyasato T, Nakatsu R , 1998 Bimodal emotion recognition by man and machine, in: proc. Workshop on Virtual Communication Environments, 1998
- Jaimes A. and Sebe N. 2007, Multimodal human–computer interaction: A survey, in: Computer Vision and Image Understanding, 2007
- Jaimes A. and Sebe N., 2005, Multimodal human computer interaction: a survey, IEEE International Workshop on HCI, 2005
- Kim, Jonghwa, 2004, Sensing Physiological Information, Applied Computer Science, Workshop Santorini, HUMAINE WP4/SG3, 2004, online available on <http://emotion-research.net/>
- Lisetti C.L. and Nasoz F. 2004, Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals, in Journal on Applied Signal Processing, Issue11, 2004
- Maat L. and Pantic M., 2006 Gaze-X: adaptive affective multimodal interface for single-user office scenarios, in Proceedings of the 8th international conference on Multimodal interfaces, 2006
- Marcel S., 2002, Gestures for multi-modal interfaces: a review, Technical Report IDIAP-RR 02-34, 2002.
- Nakasone A., Prendinger H., Ishizuka M. 2005, Emotion Recognition from Electromyography and Skin Conductance, in proc of The Fifth International Workshop on Biosignal Interpretation, 2005 - Citeseer
- Nasoz F., Alvarez K, Lisetti C. L. and Finkelstein N. Emotion recognition from physiological signals using wireless sensors for presence technologies; Cognition, Technology & Work; Volume 6, Number 1 / February, 2004
- Oudeyer Pierre-Yves, 2003, The production and recognition of emotions in speech: features and algorithms International Journal of Human-Computer Studies 59, 2003
- Scherer, Klaus R., Rainer Banse, and Harald G. Wallbott. 2001. Emotion Inferences from Vocal Expression Correlate across Languages and Cultures, Journal of Cross-Cultural Psychology 32/1, 2001
- Schröder M, Cowie R, Douglas-Cowie E, Westerdijk M. Gielen & S., 2001, Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis, In: Proceedings Eurospeech, 2001
- Schröder, Marc, 2004, Dimensional Emotion Representation as a Basis for Speech Synthesis with Non-extreme Emotions, Publisher: Springer Berlin / Heidelberg

- Schuller B., Lang M., Rigoll G, 2002, Multimodal emotion recognition in audiovisual communication, proc. of IEEE International Conference on Multimedia, 2002
- Sebe N., Cohen I. , Gevers T. , Huang T.S., 2006, Emotion Recognition Based on Joint Visual and Audio Cues, in: Proc. International Conference on Pattern Recognition (2006),
- Sebe N., Cohen I., Gevers T., and Huang T., 2005, Multimodal Approaches for Emotion Recognition: A Survey, in: Proceedings of SPIE, 2005
- Slavova, V., Verhelst, W., Sahli H. 2008, A cognitive science reasoning in recognition of emotions in audio-visual speech, in: International Journal of Information Theories and Applications, vol. 15 / 2008,
- Takahashi, Kazuhiko 2004; Remarks on Emotion Recognition from Bio-Potential Signals; in proc of 2nd International Conference on Autonomous Robots and Agents, 2004 Palmerston North, New Zealand
- Takarishi K. 2004, Comparison of Emotion Recognition Methods from Bio-Potential Signals, in Japanese Journal of Ergonomics, vol. 40 n2, 2004.
- Turk M., 2001, Gesture recognition, Handbook of Virtual Environment Technology, K. Stanney (Ed.), 2001.
- Wan V., Ordelman R., Moore J., Muller R., (2005) "AMI Deliverable Report, Describing emotions in meetings", internal project report, On line available <http://www.amiproject.org/>
- Zeng Z., Pantic, M., Roisman, G.I., Huang, T.S., 2009, A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions, in: Pattern Analysis and Machine Intelligence, 1, 2009
- Zeng Z., J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T.S. Huang, D. Roth, S. Levinson, 2004, Bimodal HCI-related affect recognition, ICMI, 2004.
- Zeng Z., Jilin Tu, Ming Liu, Thomas S. Huang, Brian Pianfetti, Dan Roth, and Stephen Levinson 2007, Audio-Visual Affect Recognition, IEEE Transactions on multimedia, V9, N2, 2007

---

### Authors' Information

---

*Velina Slavova* - New Bulgarian University, Department of Computer Science; e mail: [vslavova@nbu.bg](mailto:vslavova@nbu.bg)

*Hichem Sahli* - Vrije Universiteit Brussel, Department of Electronics & Informatics; e mail: [hsahli@etro.vub.ac.be](mailto:hsahli@etro.vub.ac.be)

*Werner Verhelst* - Vrije Universiteit Brussel, Department of Electronics & Informatics; [wverhels@etro.vub.ac.be](mailto:wverhels@etro.vub.ac.be)

---

## PROGNOSTICATION OF EFFICIENCY OF MEDICAL AND PROPHYLACTIC MEASURES AT DIFFERENT HOMOEOSTASIS VIOLATION OF HUMAN ORGANISM BY MARKOV PROCESSES THEORY

Boris Samura, Anatoly Povoroznuk, Olga Kozina, Elena Visotskaja,  
Elena Chernykh, Nikolay Shukin, Andrei Porvan

*Abstract:* This article is devoted to the questions of prognostication of efficiency medical and prophylactic measures at renewal of the broken equilibrium of human organism by vegetable medications. The methods of modeling of different processes are considered in the article from position of systems analysis. There was chosen Markov processes theory for description of renewal of the broken homoeostasis process, which rotined high correlation between clinical supervisions and forecast process.

*Keywords:* casual process, Markov chains, medical and prophylactic measures, MAIS, MFPS.

*ACM Classification Keywords:* I.6 Simulation and modeling, I.6.3 Applications; J.3 Life and medical sciences – Medical information systems

---

### Introduction

For today increased interest of doctors is concentrated in the alternative methods of human treatment and, especially, by phitotherapy. It is method of human organism treatment by medications of phitomaterials, which can be used as independent or additional type of treatment and prophylaxis of different diseases and rehabilitation of patients with the chronic diseases. A new coil in development of this type of treatment demanded the revision of attitude toward it, and development of modern methods of lead through and estimation of it efficiency. Phitotherapy foresees setting of medical plants, allowing individualizing the process of treatment taking into account classification of illnesses, their etiologic nosotropic essence, and to receives most and on possibility fast clinical effect with the minimum sides' actions, that is not always arrived at the use of synthetic medicinal preparations.

One of important stages of prophylaxis and treatment of different diseases is the ground of application expedience of one or another type of therapy. As result, the research directed on the exposure of degree of influencing of medications of vegetable origin on a time of the broken renewal of human organism homoeostasis is great scientific and practical interest.

---

### Set the problem and analysis of existent works

In connection with introduction of mathematic and computer technologies to human knowledge field the last years interest of doctors increased by the different mathematical methods, approaches and estimation facilities of the human organism state at the different methods of therapy, in particular estimations of change of the human organism state at treatment by phito-medications (PHM) [Camypa, 2003]. Application of mathematical methods and facilities allows to the specialist to obtain complete information about processes what is going on in the human organism and in good time to produces correction of the appointed treatment.

---

Problem related to estimation of process of renewal of the broken equilibrium of human organism continues to remain actual, because the effective methods of its decision not are found thus far. Presently a systems research receives wide distribution in not only biology and medicine but also pharmacology. Efficiency of the appointed treatment in a great deal depends on quality of application of systems approach, not having an alternative in the conditions of scientific prognostication and medically-practical activity in the new problem situation related to the phitotherapeutic method of complex individual renewal of the broken equilibrium of human organism.

One of main instruments of system analysis of any explored process is its mathematical modelling. In obedience to the base classification of methods of modelling of the systems and processes, offered by F.E. Temnikov, select two classes:

- methods and models directed on activating of intuition and experience of specialists (MAIS);
- methods and models of formalization presentation of the systems (MFPS).

Such division of methods is in accordance with the basic idea of systems analysis, which consists of combination in models and methods of formal and informal presentations, that helps in the choice of methods of formalization and analysis of problem situation [Волкова, 1993, Славич, 1989].

Therefore, MAIS imply verbal description of problem situation at which no necessity is in proof of the explored process. In the turn of MFPS allow to transform verbal description in formal one. This process is important component part of decision-making process.

The methods of formalization presentation of the systems can be divided to four classes:

- analytical methods;
- methods of discrete mathematics;
- statistical methods;
- methods of graphic presentation.

Graphic presentation is the comfortable mean of research of structures and processes in the complex systems (for example, human organism), and mean of organization of human and hardware co-operation. For description of different processes, this class of methods is used by the directed count of the states or network graph. However much application of these methods at estimation of the organism state, by virtue of the specificity, does not allow considering feedbacks and cyclic processes arising up in a process treatment of human.

The methods of discrete mathematics make theoretical basis of development of simulation and informative and searching languages and include set-theoretic, logical, linguistic and semiotics methods and models. A set-theoretic method are based on the concepts of great number, relations of great numbers and continuum and is used as summarizing a language at comparison of mathematics and other disciplines. The common state of organism, described by these methods, can be represented by the aggregate of great numbers and subsets of heterogeneous chambers with the arbitrarily entered elements and relations. However, introduction of arbitrary relations results in that in formalization description of the organism state can will be revealed an insoluble contradiction that does not allow operating by the received set-theoretic model the same way, as with classic analytical or statistical correlations, guarantying authenticity of the got results. The methods of mathematical linguistics and semiotics are a comfortable vehicle for formalization of decisions in tasks with a great initial vagueness and are one of constituents of construction of the complex systems of decision-making. However applying these methods it is necessary to have because of, that at complication of formal model by the rules of arbitrary grammar and semiotics it is hardness enough to guarantee authenticity of the got results, and



---

explanation of the got results not always carries objective character. The methods of mathematical logic are used at research of the systems of different nature, in which character of interrelations between elements not is clear and their analytical presentation is not possible, and statistical researches did not result in the exposure of statistical conformities to the law. However needed it is to have because of, that description of the systems by the methods of mathematical logic implies the use of logical base statutory Boolean algebra. In also time semantic possibility of logical methods is strictly limited, and the use of multiple-valued a logician appears enough by labour intensive procedure and requires the special proofs of authenticity of the received results [Гилл, 1985].

The widest distribution was got by analytical methods. These methods allow adequately representing the explored process by the determined values or dependences. These methods are widely used in the optimizations tasks of planning and analysis of the different systems in biology and medicine. However, the use of these methods requires the obvious pointing of all connections between components and goals of the system as analytical dependence.

The human organism is the open system which different external factors influence on. The processes what is going on in an organism carry probabilistic character partly. That is why it is possible to assume that the change of homoeostasis of organism also will carry probabilistic character. As result application of all higher described classes of methods and models for adequate description of homoeostasis change process as a result of complex therapy vegetable medications are not possible.

Application of the known methods of the statistical programming allows investigating the processes of the studied system without the exposure of the clear determined conformities to the law. It is explained to those, that at application of statistical presentations the process of raising of task is partly replaced by statistical researches or expert estimations, the result of which with the certain share of probability influences on the conduct of all system on the whole.

At the modelling of the complex system, a few models are usually used from a number the varieties mentioned above. Any system can be represented by different ways, which differ from each other on complication and working out in detail. In such situations for the generalized description of work of the modelling process it can be applied the imitation modelling.

All simulations models are models as the so-called black box. It means that it provide delivery of output signal of the system, if an entrance signal acts on its influencing subsystems. Therefore, for the receipt of necessary results it is necessary to carry out driving away of simulations models.

A simulation model is the special programmatic complex, which allows imitating activity of some complex object and reflects the great number of parameters subject to time-history and space.

On the stage of research and analysis of the systems at construction and realization of different models, the method of statistical design of Monte Carlo, which is based on the use of casual values, is widely used. Essence of method of statistical modelling is taken by construction for the process of functioning of the explored system of some modelling algorithm imitating the conduct and co-operation of elements of the system taking into account accidental influences and influences of external environment. The series of particular values of the sought after values or functions statistical treatment of which allows to receives information about the conduct of the real object or process in the arbitrary moment of time turn out as a result of statistical design of the system. However, for achievement of sufficient statistical stability of the system and reliable exactness of description of the explored process plenty of experiments of the system are needed. In addition, the use of method of Monte Carlo allows

building a «fictitious» model only, not having connection with an object or process of design. Markov models, in-use for the analysis and synthesis of calculable structures, which can be considered as stochastic systems without the consequences, allow taking into account this failing [Панкратова, 2005].

---

### Basic material and research results

---

Markov chain models are a natural approach to take when modeling the transitions of patients between discrete health states over time, for example, the progression over stages of a disease. There is a distinction between discrete-time Markov chains, whereby we consider transitions to occur at fixed points in time and we work with transition probabilities, and continuous time Markov chains, whereby transitions can occur at any point in time and we work with transition rates. The use of discrete-time Markov chain models in medical decision applications dates back to the work of others have championed the use of continuous-time Markov chain models based on stochastic trees and have pointed out the mathematical convenience and simplicity of these models as long as the same rates are applied over a full lifetime. A discrete-time model can approximate a continuous-time Markov model by defining a cycle length of interest (for example, yearly or 6-month intervals). The advantage of the forward equations approach is that transition rates, rather than transition probabilities, form a common basis for combining information from different studies.

The disadvantage of Markov models is that departure from the simple assumptions of stationary, first-order Markov chains while, conceptually possible, makes for disproportionate degrees of difficulty in analysis and computation. Moreover, like all other succession models, the validation of Markov models depends on predictions of system behavior over time, is therefore frequently difficult, and may even be impossible, for long periods [Медик, 2007, Токмачев, 2003].

Renewal of the broken equilibrium of organism of human with the help PHM it is possible to represent as casual process of transition from one state in other, at which each of the states describes the degree of violations passing in an organism.

However, there is some system of  $E$ , describing the process of renewal of the broken equilibrium of organism, which in the process of functioning can adopt the different states  $E_i, i = \overline{1, K}$  with probabilities  $P_i(t)$ .

It is known that the human organism state in the moment of time can be classified on the degree of disease as:  $E_1$  as the conditionally healthy;  $E_2$  as the initial changes;  $E_3$  as the easy degree;  $E_4$  as the middle degree;  $E_5$  as the heavy degree;  $E_6$  as the extremely heavy degree;  $E_7$  as the fatal outcome.

For research of influencing of synthetic and phitotherapeutic preparations on the renewal process of the broken equilibrium of human organism 150 patients were inspected. Coming statistical data from the vectors of apriory probabilities of finding were received in  $i^{\text{th}}$  state of organism of patient –  $P_i(0)$ .

$$P_1(0) = [0.985; 0.015; 0.00; 0.00; 0.00; 0.00; 0.00];$$

$$P_2(0) = [0.03; 0.955; 0.015; 0.00; 0.00; 0.00; 0.00];$$

$$P_3(0) = [0.00; 0.03; 0.94; 0.03; 0.00; 0.00; 0.00];$$

$$P_4(0) = [0.00; 0.00; 0.015; 0.97; 0.15; 0.00; 0.00];$$

$$P_5(0) = [0.00; 0.00; 0.00; 0.045; 0.955; 0.00; 0.00];$$

$$P_6(0) = [0.00; 0.00; 0.00; 0.00; 0.015; 0.985; 0.000];$$

$$P_7(0) = [0.00; 0.00; 0.00; 0.00; 0.00; 0.00; 1.00].$$

A finite Markov chain is a process with a finite number of states in which the probability of being in a particular state at step  $(i+1)$  depends only on the state occupied at step  $i$ .

The dynamic supervision after patients included the clinical, laboratory-biochemical and bacteriologic examinations. After rising of diagnosis and determination of values of the initial states setting of treatment was conducted. Controls examinations were conducted during a semi year first time per a month.

At every the instant the system can is in one of seven states. Except for the initial state of the system (in the initial moment of the time  $t=0$ ), being in which a patient appealed to the doctor, known also one step transition

$P_{mn} = P\{E_i = E_n | E_{i-1} = E_m\}$ ,  $m, n = 1..7$  (see fig.1), where one step is equal to the interval between the moments of conducting of control researches. Consequently, the random process of state transition of the organism  $E_i = E(t)$  is the homogeneous Markovian chain.

$$P_{ij}^{(1)} = \begin{pmatrix} 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.93 & 0.06 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.90 & 0.07 & 0.03 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.83 & 0.14 & 0.03 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.80 & 0.13 & 0.07 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.67 & 0.32 & 0.01 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{pmatrix}$$

Fig. 1 Matrix of transitional probabilities from one state in other for the explored group

Transition probabilities from one state in other can be presented in the following as (fig. 2):

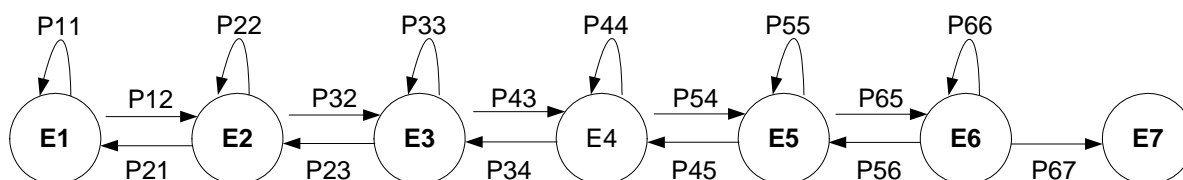


Fig. 2 Marked graph of the states for the model

We will take on to every no zeroing element of the transitions probabilities matrix during one-step  $P_{mn}$  the random variable  $T_{nm}$  with the function of distributing  $F_{nm}(t_{nm} \leq t)$ . For the examined system the random variable  $T_{nm}$  is discrete delay time of patient in a state of  $E_n$ , on condition that following state will be  $E_m$ . In other words, a patient remains in a state  $E_n$  in the flow time  $T_{nm}$ , before his state will be replaced on  $E_m$ . After the  $E_m$  state achievement «instantly» (in accordance with matrix of transitional probabilities), a next state  $E_k$  is selected. Since the  $E_k$  state is chosen delay time in  $E_k$   $T_{kl}$  relies equal with function of probabilities distribution  $F_{kl}(t)$ . This process can proceed long time, choosing an each time independent next state and delay time.

It ensues from the resulted determination, that if to ignore random character of delay time and to take interest only transition moment, a process  $E(t)$  will be a Markov's homogeneous chain (or embedded chain).

However at the account of stay of process  $E(t)$  in the different states during the random intervals of time will not satisfy to equalization Markov (if not all times are up diffused exponentially). Consequently, a process is Markov's process only in the transition moments. The said justifies the name of «semi -Markov process» or «semi-Markov chain».

At the given initial state the further conduct of semi-Markov process fully concerns by the matrix of transitional probabilities  $\{P_{ij}(t)\}$ ,  $i, j = 1..7$ , and matrix of functions of probabilities distributions  $\{F_{ij}(t)\}$ . We will assume that in some moment of time, taken as a start ( $t_0=0$ ), the system is in a state of  $E_i$ . Let the following state is chosen (with probability of  $P_{ij}$ ) the state of  $E_j$ . Then by theorems of product and adding probabilities, we find the absolute function of distributing of complete delay time in a state  $E_i$

$$F_i(t) = P(t_i < t) = \sum_{j=1}^7 P_{ij}^{(1)} \cdot F_{ij}(t), \quad j=1..7 \quad (1)$$

Middle value of absolute delay time  $T_i$  in state  $E_i$  equal

$$T_i = \sum_{j=1}^7 P_{ij}^{(1)} \cdot \bar{T}_{ij} \quad (2)$$

We will pass now to the decision of main task arising up at the analysis of semi-Markov processes, i.e. to the calculation of probabilities of the states. Let  $C_{ij}(t)$  is conditional probability at which in the time moment  $t$  system is in a state of  $E_j$ , if in the time moment  $t_0=0$  it was in a state of  $E_i$ . Probability of  $C_{ij}(t)$  can be named interval-transitional probability. The system, starting from the initial state of  $E_i$ , can get in the state of  $E_i$  in the time moment equal  $t$  by different ways. Firstly, if  $E_i = E_j$  it can stay in state  $E_i$  during all interval  $t$  or, going out from the state  $E_i$  at least once, it all the same goes back into the state  $E_j = E_i$  till the time  $t$ . Secondly, the system can get in the arbitrary state of  $E_j$ , occupying in the time moment  $\tau$  some transient state of  $E_k$ . Probabilities of these two mutually eliminating possibilities must be added:

$$C_{ij}(t) = \delta_{ij} \cdot \left( 1 - \sum_{j=1}^7 P_{ij}^{(1)} \cdot F_{ij}(t) \right) + \sum_{k=1}^7 P_{ik}^{(1)} \cdot \sum_{\tau=0}^t F_{ik}(\tau) \cdot C_{kj}(t-\tau), \quad 1 \leq i, j \leq 7, \quad (3)$$

where  $\delta_{ij}$  – Kronecker's symbol,  $\delta_{ij} = 1$  at  $i = j$  and  $\delta_{ij} = 0$  at  $i \neq j$ .

The first member on the right in equation (3) takes into account probability of event  $E_i = E_j$ , because  $(1 - \sum_{j=1}^7 P_{ij}^{(1)} \cdot F_{ij}(t))$  is probability at which the system will not leave the state  $E_i$  during interval time  $t$ . The second

member in (3) presents probability of sequence of events, while the system accomplishes transition from  $E_i$  to  $E_k$  (it can be even in itself) to moment  $\tau$  and then passes from the state of  $E_k$  to the state of  $E_j$  for remaining time of  $t - \tau$ . Probabilities of every possible transition are added on all transient states of  $E_k$ , in which transitions are possible from the initial state of  $E_i$ , and are summing on various times of transition  $\tau$  of  $\tau$  between 0 and  $t$ .

System of linear integral equations (3) is basis. It gives expression of interval-transitional probabilities by main characteristics of semi-Markov process. Nevertheless, to get the analytical decision of this system not simply. Some simplification gives application of method of Laplace's transformation.

For simplification of calculations according (3) and when probabilities distributions of random processes  $F_{ij}(t)$  for every state  $E_i$  unknown, we will ignore random character of delay time in every state and to take interest only transition moments. In other words, let assume that we have Markov's homogeneous chain for which probability of staying in the state  $E_k$ , in the moment of time of  $(t + \Delta t)$  will concern by the formula of complete probability:

$$P_K(t + \Delta t) = P_1(t) \cdot P_{1K} + P_2(t) \cdot P_{2K} + \dots + P_K(t) \cdot P_{KK} + \dots + P_n(t) \cdot P_{nK}. \quad (4)$$

with condition that  $\sum_{i=1}^7 P_i = 1$ . According to (4) with given vectors of apriory probabilities of staying in every states

$P_i(0)$  and one-step transition matrix  $P_{ij}^{(1)}$  we we received sets of state probabilities during first 12 months.

In tabl.1 shown that if patient first time come to doctor with middle degree of disease (i.e.  $E_4$  state exist at  $t=0$  with probability  $P_4(t=0)=0.75$ ) then on 3step or after 3 months probability of state  $E_1$  will equal 0.653 to contrary with probability of state  $E_4$  which will equal 0.029. Thus in tabl.1 we analyze number of step at which probability of state  $E_1$  – the conditionally healthy – higher than any other probabilities as dependence on initial vectors  $P_i(0)$ . For all initial vectors  $P_i(0)$  in all next steps after steps marked into tabl.1, probability of state  $E_1$  is increase and probabilities of others states are decrease.

The Markov chain on fig.1 is absorbing because it has two absorbing states, and from every state it is possible to go to absorbing states. As known, in a finite number of steps the chain will enter an absorbing state and then stay in that state. Also, the long-term trend depends on the initial state – changing the initial state can change the final result. This property of our analyzed system is reflected into tabl.1 and in common distinguishes absorbing Markov chain from regular Markov chains, where the result is independent of the initial state.

Table 1. Probabilities of states  $E_i$  for different initial vectors  $P_i(t=0)$

Step, t	$P_1(t)$	$P_2(t)$	$P_3(t)$	$P_4(t)$	$P_5(t)$	$P_6(t)$	$P_7(t)$
0	0.985	0.015	0.000	0.000	0.000	0.000	0.000
1	0.999	0.001	0.000	0.000	0.000	0.000	0.000
0	0.030	0.955	0.015	0.000	0.000	0.000	0.000
2	0.993	0.006	0.001	0.000	0.000	0.000	0.000
0	0.000	0.030	0.940	0.030	0.000	0.000	0.000
2	0.816	0.133	0.042	0.008	0.001	0.000	0.000
0	0.000	0.000	0.140	0.750	0.110	0.000	0.000
3	0.653	0.223	0.084	0.029	0.008	0.003	0.000
0	0.000	0.000	0.000	0.045	0.955	0.000	0.000
4	0.571	0.232	0.117	0.053	0.019	0.007	0.001
0	0.000	0.000	0.000	0.000	0.015	0.985	0.000
6	0.644	0.174	0.093	0.048	0.018	0.008	0.016

To obtain information about the time to absorption in an absorbing Markov chain, we first calculate the fundamental matrix  $B_m$  for our patients from one-step transition matrix  $P_{ij}^{(1)}$  [Гармоткина, 2005]. Let  $I_1$  represent the  $1 \times 1$  identity matrix in the upper left corner of  $P_{ij}^{(1)}$ ; let  $O$  represent the matrix of zeros in the upper right; the  $R$  represent the matrix in the lower left; and let  $Q$  represent the matrix in the lower right. Using these symbols,  $P_{ij}^{(1)}$  can be written as

$$P_{ij}^{(1)} = \begin{bmatrix} I_1 & O \\ R & Q \end{bmatrix}.$$

The fundamental matrix for the absorbing Markov chain is matrix  $B$ , where

$$B = (I_n - Q)^{-1}. \quad (5)$$

Here  $I_n$  is the  $n \times n$  identity matrix corresponding in size to matrix  $Q$ , so that the difference  $I_n - Q$  exist. For our Markov chain transition matrix  $P_{ij}^{(1)}$  can be rewritten in follow view:

$$P_{ij}^{(1)} = \left[ \begin{array}{cc|cccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0.93 & 0.07 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.90 & 0.07 & 0.03 & 0 & 0 \\ 0 & 0 & 0 & 0.83 & 0.14 & 0.03 & 0 \\ 0 & 0 & 0 & 0 & 0.80 & 0.13 & 0.07 \\ 0 & 0.01 & 0 & 0 & 0 & 0.67 & 0.32 \end{array} \right]$$

Thus, our fundamental matrix will be

$$B = \left[ \begin{array}{ccccc} 1.075 & 0 & 0 & 0 & 0 \\ 1.075 & 1.111 & 0.040 & 0.002 & 0.000 \\ 1.075 & 1.111 & 1.245 & 0.047 & 0.005 \\ 1.073 & 1.109 & 1.243 & 1.295 & 0.133 \\ 1.058 & 1.093 & 1.225 & 1.276 & 1.602 \end{array} \right]$$

Let our Markov chain currently in state  $i$ . The expected number of times that the chain visits state  $j$  at this time is 1 for  $i$  and 0 for all other states. The expected number of times that the chain visit state  $j$  at next step is given by element in row  $i$ , column  $j$  of the transition matrix  $Q$ . The expected number of times the chain visits state  $j$  two steps from now is given by corresponding entry in matrix  $Q^2$ . The expected number of visits in all steps is given by  $I_n + Q + Q^2 + Q^3 + Q^4 + \dots$ . To find out whether this infinite sum is the same as  $(I_n - Q)^{-1}$ , multiply the sum by  $(I_n - Q)$ :

$$(I + Q + Q^2 + Q^3 + \dots) \cdot (I - Q) = I + Q + Q^2 + Q^3 + \dots - Q - Q^2 - Q^3 + \dots = I,$$

which verifies our results.

It can be shown that

$$P_{ij}^{(k)} = \left[ \begin{array}{c|c} I_n & O \\ \hline (I + Q + Q^2 + \dots + Q^{k-1}) \cdot R & Q^k \end{array} \right],$$

where  $I_n$  is the  $n \times n$  identity matrix. As  $k \rightarrow \infty$ ,  $Q^k \rightarrow O_n$ , the  $m \times m$  zero matrix, and

$$P_{ij}^{(k)} = \left[ \begin{array}{c|c} I_m & O \\ \hline B \cdot R & O_n \end{array} \right],$$

so  $BR$  gives the probabilities that if the system was originally in a non-absorbing state, it ends up in one of absorbing states.

The entry  $b_{ij}$  of  $B$  gives the expected number of times that the process is in the transient state  $E_j$  if it is started in the transient state  $E_i$ .

The total number of steps expected before absorption equals the total number of visits which expected to make to all the non-absorbing states. This is the sum of all the entries in the  $i^{\text{th}}$  row of  $B$ . For our fundamental matrix the total number of steps expected before absorption (means state  $E_1$ ) show in tabl.2.

Table 2. The total number of steps expected before absorption for different start state

Start state	$E_2$	$E_3$	$E_4$	$E_5$	$E_6$
№ steps	1.075	2.223	3.483	4.855	6.254

Comparing of data from tabl.1 and tabl.2 show aproximally same results.

---

## Conclusion

---

Theory of the Markov processes allows numeral to describe the behavior of the complex system, namely influence of phito-medications facilities on human organism. Thus, it is possible to get a prognosis on duration of stay of patient on all selected states of medically diagnostic cycle depending on prevalence's. The numerical results of mathematical modeling with the use of simplification through assumption that analyzed process can be described via embedded Markov chain, shows high correlation to clinical observations. However, mathematically will be more correctly to use theory of semi-Markov processes in our case. As shown, semi-Markov processes theory cannot give finite analytical results. But if clinicians will give distributions of probabilities of delay time in every state during the treatment it may be to use simulation to receive numerical results of probabilities that patient will be on the state in any time.

---

## Bibliography

---

- [Самура, 2003] Б.А. Самура, В.Ф. Черных, И.П. Банный. Фитотерапия в клинике внутренних болезней. Х.: Изд-во НФаУ: Золотые страницы, 2003.
- [Волкова, 1993] В.Н. Волкова, А.А. Денисов, Ф.Е. Темников. Методы формализованного представления систем. СПб.: СПбГТУ, 1993.
- [Гилл, 1985] Ф. Гилл, Н. Мюррей, М. Райт. Практическая оптимизация. М.: Мир, 1985.
- [Славиц, 1989] М.Б. Славиц. Методы системного анализа в медицинских исследованиях. М.: Медицина, 1989.
- [Москинова, 2000] Г.И. Москинова. Дискретная математика для инженеров. М.: Логос, 2000.
- [Горбачев, 2005] В.А. Горбачев. Технологии моделирования систем. Х.: «Компания СМИТ», 2005.
- [Панкратова, 2005] Н.Д. Панкратова. Системный анализ: проблемы, методология, приложения. К.: Наукова думка, 2005.
- [Медик, 2007] В.А. Медик, М.С. Токмачев. Математическая статистика в медицине. М.: Финансы и статистика, 2007.
- [Токмачев, 2003] М.С. Токмачев. Цепи Маркова в прогнозировании медико-социальных показателей. Обзорение прикладной и промышленной математики. Т.10. Вып. 2., 2003.
- [Гармоткина, 2005] О.В. Гармоткина. Марковская модель заболевания населения. Искусственный интеллект. Вып. 2., 2005.

---

**Authors' Information**

---

*Boris Samura* – doctor of pharmaceutical sciences, professor, honoured worker of scitech Ukraine, academician of ANTK of Ukraine, Head of Pharmacotherapy Department of National University of Pharmacy, Darwin street, 8/10, Kharkov, 61002, Ukraine

*Anatoly Povoroznuk* – PhD, professor of Computers and Programing Department of National Technical University 'KPI', Frunze street, 21, Kharkov, 61002, Ukraine

*Olga Kozina* – PhD, lecturer of Computers and Programing Department of National Technical University 'KPI', Frunze street, 21, Kharkov, 61002, Ukraine; [okaraban@rambler.ru](mailto:okaraban@rambler.ru)

*Elena Visotskaja* – PhD, lecturer of Biomedical Electronic Devices and Systems Department of Kharkov National University of Radio Electronics, Lenina Av., 14, Kharkov, 61166, Ukraine; e-mail: [diagnost@kture.kharkov.ua](mailto:diagnost@kture.kharkov.ua)

*Elena Chernykh* – PhD, lecturer of Computers and Programing Department of National Technical University 'KPI', Frunze street, 21, Kharkov, 61002, Ukraine

*Nikolay Shukin* – probationer is a researcher of Biomedical Electronic Devices and Systems Department of Kharkov National University of Radio Electronics, Lenina Av., 14, Kharkov, 61166, Ukraine

*Andrei Porvan* – engineer of Biomedical Electronic Devices and Systems Department of Kharkov National University of Radio Electronics, Lenina Av., 14, Kharkov, 61166, Ukraine; e-mail: [borman\\_d@mail.ru](mailto:borman_d@mail.ru)



---

## INDIRECT SPATIAL DATA EXTRACTION FROM WEB DOCUMENTS

Dimitar Blagoev, George Totkov, Milena Staneva,  
Krassimira Ivanova, Krassimir Markov, Peter Stanchev

*Abstract:* An approach for indirect spatial data extraction by learning restricted finite state automata from web documents created using Bulgarian language are outlined in the paper. It uses heuristics to generalize initial finite-state automata that recognizes only the positive examples and nothing else into automata that recognizes as larger language as possible without extracting any non-positive examples from the training data set. The learning method, program realization and experiments are presented. The investigation is carried out in accordance and following the rules of EU INSPIRE Network.

*Keywords:* Automatic Data Extraction, Restricted Finite State Automata, Web Documents, Indirect Spatial Data, INSPIRE network.

*ACM Classification Keywords:* H.2.8 Database Applications - Data mining; F.1.1 Models of Computation – Finite State Automata

---

### Introduction

---

“*Spatial data*” is data with a *direct* or *indirect* reference to a specific location or geographic area [INSPIRE-DSM, 2007]. Our attention in this paper is given to the indirect references included in the text documents written in Bulgarian language.

The *indirect* reference to a specific location or geographic area may be of different types and formats. Because of this it is difficult to propose a common classification of such information. In the same time, one of the main characteristic of indirect references is the address, i.e. a description of the interconnection of the data with the specific location or geographic area [INSPIRE-DSAD, 2008]. Usually this is a text with common structure – location of properties based on address identifiers, usually by road name, house number, postal code, etc.

In everyday practice there are many kinds of documents containing indirect references to a specific locations or geographic areas. The kernel problem is that EU member countries use different languages and national standards for different types of indirectly given references. The automatic extraction of the references is very important for processing such documents in the INSPIRE network.

In recent years multiple machine learning approaches have been proposed for information extraction [Li et al, 2008]. A large class of entity extraction tasks can be accomplished by the use of carefully constructed regular expressions. Examples of entities amenable to such extractions include e-mail addresses, software names (web collections), credit card numbers, social security numbers (e-mail compliance), gene and protein names (bioinformatics), etc. With a few notable exceptions, there has been very little work in reducing this human effort through the use of automatic learning techniques.

In the context of information extraction, prior work has concentrated primarily on learning regular expressions over relatively small alphabet sizes and usually learning of regular expressions is done over tagged tokens produced by other text-processing steps such as POS tagging, morphological analysis, and gazetteer matching [Ciravegna, 2001].

[Rozenfeld et al, 2008] propose approach, which use the immense amount of unlabeled text in the web documents in order to create large lists of entities and relations. Based on this approach the system SRES is a self-supervised web relation extraction system that learns powerful extraction patterns from unlabeled text, using short descriptions of the target relations and their attributes.

The proposed in [Li et al, 2008] learning algorithm ReLIE takes as input not just labeled examples but also an initial regular expression, which provides a natural mechanism for a domain expert to provide domain knowledge about the structure of the entity being extracted and meaningfully restriction of the space of output regular expressions.

In 2004 a team of Prof. William Cohen from Carnegie Mellon University starts creating collection of classes for storing text, annotating text, and learning to extract entities and categorize text called MinorThird [Cohen, 2004]. It contains a number of methods for learning to extract and label spans from a document, or learning to classify spans (based on their content or context within a document). The creating of such collections is a useful tool not only for the particular investigation support, but also for creating common notion for the area as a whole.

An approach for indirect spatial information extraction by learning restricted finite state automata from marked web documents created using Bulgarian language is outlined in the paper. We use heuristics to generalize initial finite-state automata that recognizes only the positive examples and nothing else into automata that recognizes as larger language as possible without extracting any non-positive examples from the training data set.

The proposed approach is a good base for building system from the class of Semi-Automated Interactive Learning (SAIL) systems [IBM, 2009]. In the next chapters the INSPIRE network as possible practical area, the proposed approach, program realization and experiments are presented. Finally, the conclusions and steps for feature work are outlined.

---

## The INSPIRE Network

---

Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 for establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), was published in the Official Journal of the European Union on 25 April 2007 and was entered into force on 15 May 2007 [INSPIRE Directive, 2007]. The main goal of the Directive is to establish a new common approach for processing the spatial data in all EU member countries.

A simplified view to the processing of data today is shown in the Figure 1. In most cases, each EU member state uses input data according to different, often undocumented or ill-documented data specifications and uses different methods to process the input data to produce more or less similar information relevant for policies within the Community [INSPIRE-DSM, 2007]. For instance two different states "A" and "B" have their own specific data specifications and data sets and their own processing methods.

The methodology described in the Directive aims for a better understanding of the common user requirements for data in INSPIRE. It focuses on the development of harmonized data specifications for the input data. This way all input data from the different member states will follow the same data specifications and the same processing steps to derive the information. The input data in the member states and their maintenance procedures will typically be more-or-less the same prior to INSPIRE, but in addition the data will be provided by the network services of the member states following the harmonized data specifications [INSPIRE-DSM, 2007]. The updated schema based on the proposed methodology for two states "A" and "B" is shown on Figure 2.

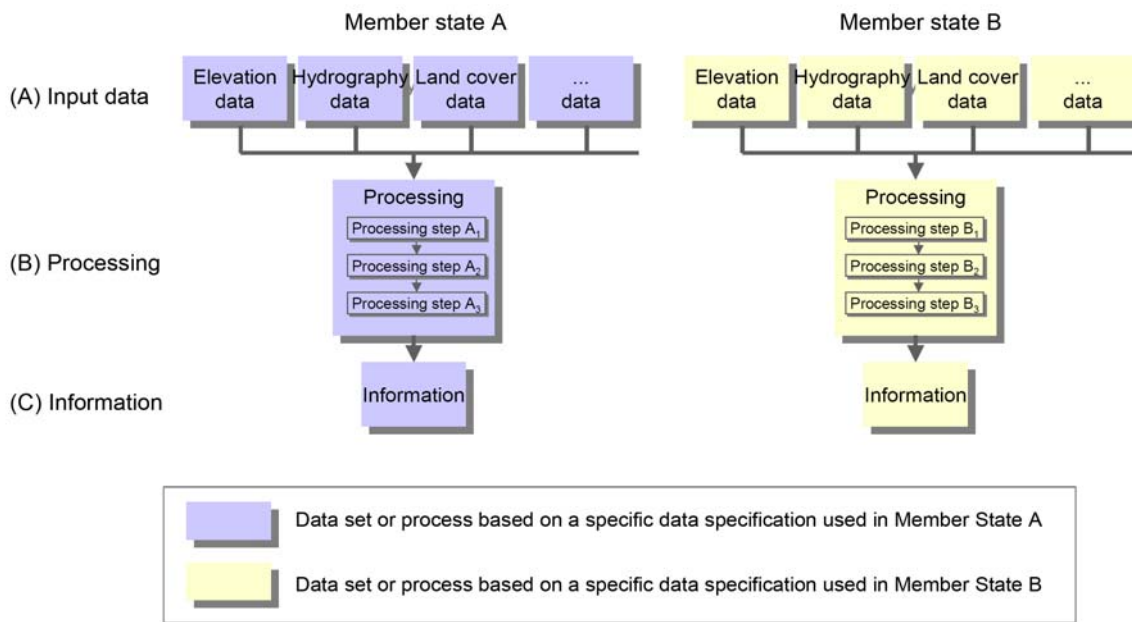


Figure 1. Current situation is "Data stovepipes"

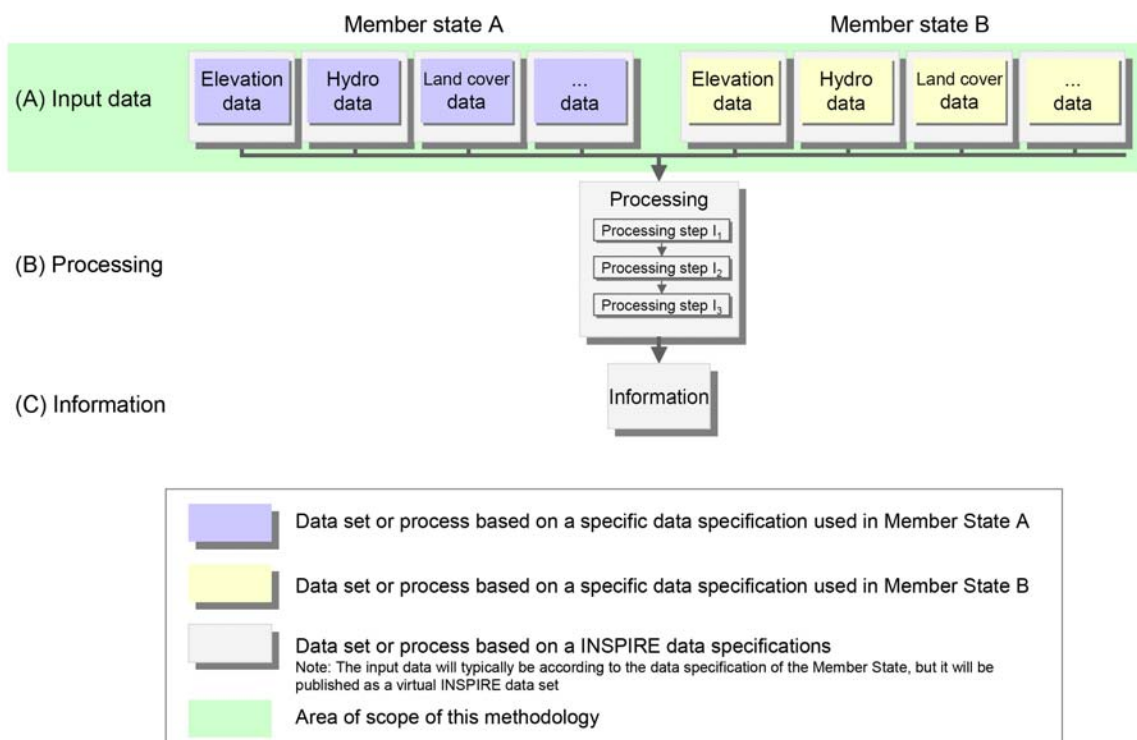


Figure 2. Target situation: Harmonized data views, eliminating data stovepipes

INSPIRE should be based on the infrastructures for spatial information that are created by the member states and are designed to ensure that:

- spatial data are stored, made available and maintained at the most appropriate level;

- 
- it is possible to combine spatial data from different sources across the European Community in a consistent way and share them between several users and applications;
  - it is possible for spatial data collected at one level of public authority to be shared between other public authorities;
  - spatial data are made available under conditions which do not unduly restrict their extensive use;
  - it is easy to discover available spatial data, to evaluate their suitability for the purpose and to know the conditions applicable to their use.

For these reasons, the Directive focuses in particular on five key areas:

- metadata;
- the interoperability and harmonization of spatial data and services for selected themes (as described in Annexes I, II, III of the [INSPIRE Directive,2007]);
- network services and technologies;
- measures on sharing spatial data and services;
- coordination and monitoring measures.

INSPIRE lays down the legal framework for the establishment and operation of an Infrastructure for Spatial Information in Europe. The purpose of such an infrastructure is to assist policy-making in relation to policies that may have a direct or indirect impact on the environment. *"Infrastructure for spatial information"* means metadata, spatial data sets and spatial data services; network services and technologies; agreements on sharing, access and use; and coordination and monitoring mechanisms, processes and procedures, established, operated or made available in accordance with the Directive [INSPIRE Directive, 2007]:

Every spatial object in a spatial data set needs to be described by a data specification specifying the semantics and the characteristics of the types of spatial objects in the data set. The spatial object types provide a classification of the spatial objects and determine among other information the properties that any spatial object may have (be they thematic, spatial, temporal, a coverage function, etc.) as well as known constraints (e.g. the coordinate reference systems that may be used in spatial data sets). This information is captured in an application schema using a conceptual schema language, which is a part of the data specification. As a result, a data specification provides the necessary information to enable and facilitate the interpretation of spatial data by an application [INSPIRE-TAO, 2007].

The logical schema of the spatial data set may and will often differ from the specification of the spatial object types in the data specification. In this case, and in the context of real-time transformation, a service will transform queries and data between the logical schema of the spatial data set and the published INSPIRE application schema on-the-fly. This transformation can be performed by the download service offering access to the data set or a separate transformation service.

*The main goal of INSPIRE is the "Interoperability"* which means the possibility for spatial data sets to be combined, and for services to interact, without repetitive manual intervention, in such a way that the result is coherent and the added value of the data sets and services is enhanced.

One important aspect of this process is the automatic extraction of spatial data and creating the corresponded metadata.

## Data Extraction by Learning Restricted Finite State Automata

The approach for indirect spatial information extraction by learning restricted finite state automata from marked web documents contains four main steps:

1. Setting up the hierarchical structure of the data to be extracted. Every element and sub-element which is to be identified has to be specified. The data structure is expressed as a tree of elements and their sub-elements.
2. Scanning and manual tagging the initial documents for the required information.
3. Extracting the examples for the different elements and building an initial parsing grammar.
4. Data extracting from new documents. The user can continue to improve the accuracy of the results by manually correcting the annotations for a particular document and add it to the learning set.

The building of the parsing grammar consists of two sub-steps:

- a) combining all positive examples;
- b) generalizing the resulting tree into restricted finite state automata.

At the sub-step a) all marked instances of the structured data are flattened in strings containing the text of the main element with special symbols marking the beginnings and ends of the sub-elements and the HTML tags in the case when the text is a web document. Then all the flattened strings from all documents are combined in a single tree. This tree is then used as the initial finite state automata. It recognizes all learned positive examples without misrecognizing negative ones.

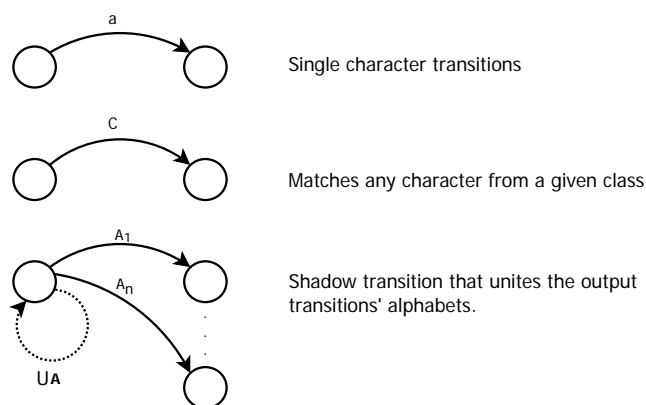


Figure 3. Elements of the restricted finite state automata [Baltes, 1992]

The sub-step b) is the generalization of the automata using heuristic methods for combining states and extrapolating the transitions' characters into predefined alphabets. After each generalization the automata is checked for consistency by re-scanning the learning texts and if the extracted data differs from the initial (manually annotated) data the modification is rolled back. There are many ways in which the finite state automata can be generalized [Baltes, 1992]. To prevent the computational complications that arise from this condition we use restricted finite state automata. The building elements that are used in these automata are (Figure 3):

- single character transition;
- matching any character from a given class;
- shadow transition that unites the output transitions' alphabets.

In addition to the automata generalization heuristics described later in the section the generalizator employs the use of a custom character class list. The class list specifies what characters belong to a given class and how many of them have to be present in a state's output transitions before class generalization is attempted. Table 1 shows one sample list which includes classes for both English and Bulgarian letters.

Table 1: Sample character class list

Min	Characters	Class
3	abcdefghijklmnopqrstuvwxyz	English lowercase
3	ABCDEFGHIJKLMNOPQRSTUVWXYZ	English capitals
3	abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ	English
3	абвгдежзийклмнопрстуфхцчшщъьюя	Bulgarian lowercase
3	АБВГДЕЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЬЮЯ	Bulgarian capitals
3	абвгдежзийклмнопрстуфхцчшщъьюяАБВГДЕЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЬЮЯ	Bulgarian
1	0123456789	Digits
1	\b \t \n \r	White space characters
1	' " " " « » ' ,	Quotes

The generalization algorithm (Figure 4) in sub-step b) is done in the following way:

1. Class generalization (top-down) which tries to generalize as much as possible output transition characters for a given state into classes;
2. State merging (bottom-up) with character comparison tries to merge a state with one of its next possible states if the two states have identical characters and classes on their output transitions. If it is successful, the two states are merged into a single state and a shadow transition is added over the union of the other output transitions' alphabets. Further testing is made to find the upper repetition limit for the newly formed state;
3. State merging (bottom-up) without character comparison, essentially same as above except it does not require two states to have comparable characters and classes on their output transitions. If it is necessary, character transitions are merged with class transitions. This operation is more prone to making erroneous generalizations or one that block the further generalization of upper states therefore it is performed after the previous generalizations;
4. Character and class merging (bottom-up) tries to merge a character transition in a given state with a class or another character transition in the same state resulting in a transition over a new class which was not predefined in the classes list;
5. State skipping (bottom-up) which tests if all output transitions on a given state can be skipped thereby advancing onto all sub-states without matching any of the transitions. Every output transition to another state is complemented with an epsilon transition (one that matches the zero-length string) to the same sub-state.

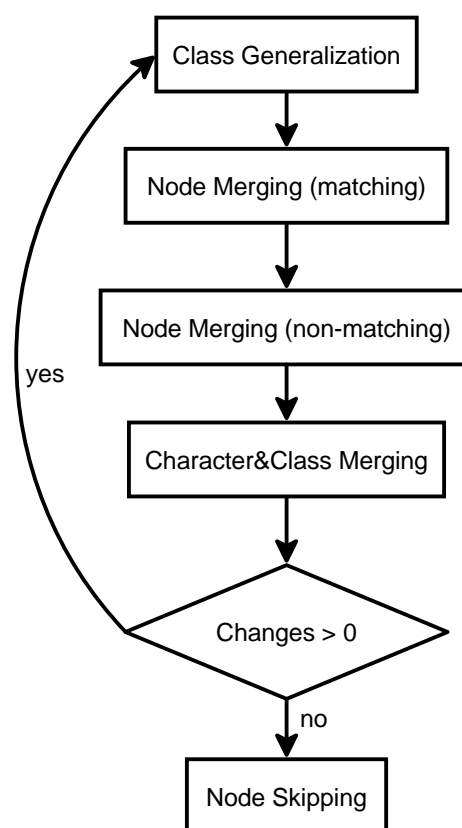


Figure 4. Generalization algorithm

After every change of the automata a test run is performed with the new automata over the learning data set. If the result differs from the initial ones the attempted transformation is rolled back. All generalization and merging

steps (without state skipping) are repeated until there are no more states which can be merged and/or generalized.

Once the module's learning phase is complete a parsing grammar is being generated. This grammar employs regular expressions for data extraction and generates structured XML output containing all elements found in the parsed documents. The sub-elements of the hierarchical data structure are encoded as named groups in the regular expressions. This grammar can then be used for performing background batch processing on a large number of documents or to analyze the produced regular expressions and make inferences for the structure of the elements of interest.

## Program Realization

The presented algorithm has been realized in the experimental system InDES. The system contains two separate modules: a graphical user interface (GUI) and a command-line learning and extracting module. The GUI is developed in C#.NET and employs the embedded Internet Explorer browser component to display the web documents. The learner and extractor are written in C++ for increased performance and smaller memory footprint. Both modules use the same html preprocessing routine for cleansing the given web documents. The cleansing's purpose is to normalize or eliminate characters in the input document without changing the structure of the contained information or the way it appears on the screen. This includes but is not limited to Unicode character normalization where explicit character codes are replaced with their respective characters and JavaScript removal (since the current version of the system does not execute JavaScript prior to learning or extracting).

A screenshot of the GUI with a loaded web document is given on Figure 5. In the left, the sub-screen for selecting the web documents contain some already connected web sites and corresponded documents. At the right the generated grammar and founded matches are shown. In the center of the screen the current document with market texts is presented.

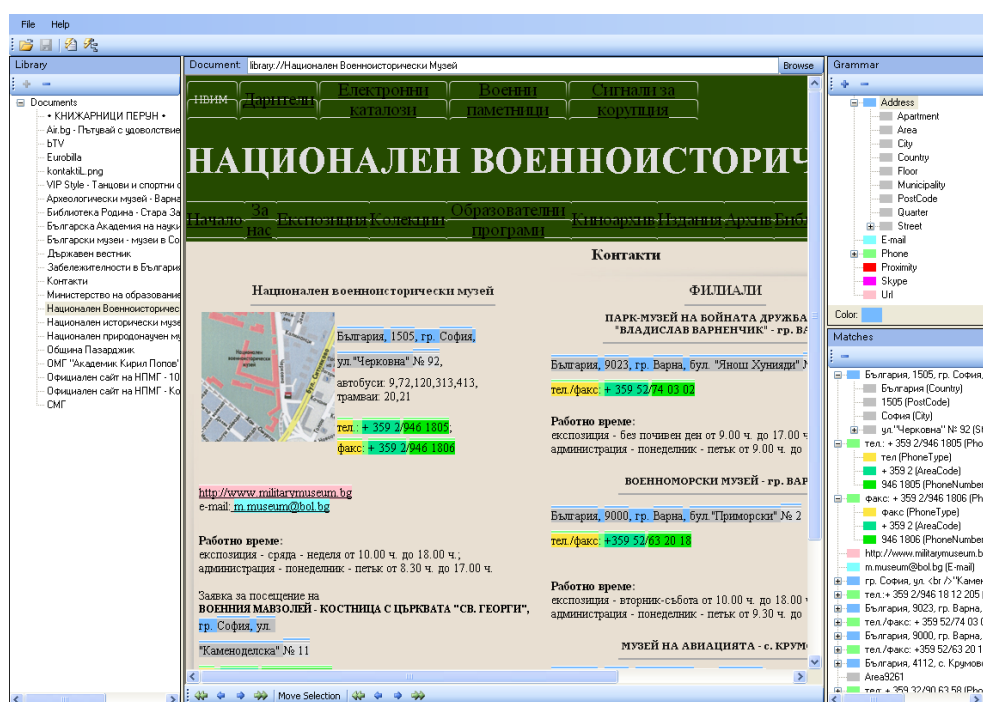


Figure 5. Screenshot of the program system InDES

## Experiments

We have made several types of experiments over different web documents in Bulgarian language for extracting elements such as addresses, phones and e-mails from randomly chosen companies' and social institutions' web pages containing contact information. To test the ability of the proposed method to extract new information we used a set of 100 pre-marked web documents in Bulgarian language for three types of elements: addresses, phones and e-mails with corresponded sub-elements. The reason for such choice is that main way for representing the indirect space information is using the addresses [INSPIRE-DSAD, 2008]. In other words, the extracting of the address is the first step for the processing the indirect space information. This information may contain different elements, represented by text sequences which are connected to any specific location or geographic area by the addresses. These elements need to be extracted, too. This means that the system need to have possibility to extract addresses as well the other types of elements from the given thematic area. To simplify the experiments, the phone numbers and e-mails are taken as such elements.

The experiments were provided following the steps of the proposed approach.

At first step the hierarchical structure of the data to be extracted was set up as it is shown on Figure 6. The structure consists of:

- addresses with sub-elements "Country", "Area", "Municipality", "City", "Post Code", "Quarter", "Street", "Floor" and "Apartment";
- phones with sub-elements "Area Code", "Phone Number" and "Phone Type";
- e-mails without sub-elements.

At the next step the data set was chosen. For the purposes of the experiments, the web document set was created using web pages for five categories organizations: companies, schools, museums, municipalities and libraries. The documents were picked out in html format using Google possibilities. For each category were selected first twenty web sites after searching for combination keywords "address" and one of keywords "company", "school", "museum", "municipality", "library" and with restriction "pages in Bulgarian".

Then, all documents from the data set was scanned and manually tagged in accordance with chosen hierarchical structure. Some of the documents are used later as instances in the learning set and other are used as instances in the testing set. At the first experiment we used ten-fold cross validation. At the second experiment the data set was split into learning set and testing set in random principle.

Since the task is to find and extract all data that represents a given element we tested the system using the following criteria:

- Recall – the percentage of manually annotated elements for which an overlapping element is found in the results of the search;

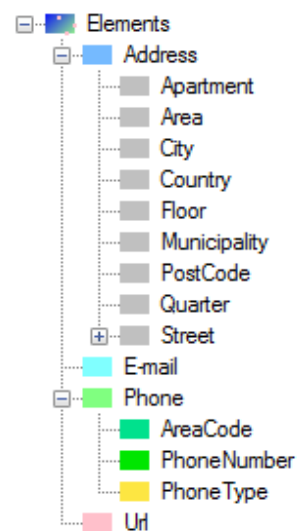


Figure 6. Sample element hierarchy for information extraction



- Precision – the percentage of found elements that overlap with a manually annotated element [Taylor, 1982];
- Accuracy – the average similarity between the original annotated elements and the correctly extracted elements.

For a similarity measure we propose to set the ratio of the length of the overlapping to the length of the union of the original and extracted texts:

$$similarity = \frac{A \cap B}{A \cup B}$$

where  $A$  and  $B$  are the original and extracted line segments respectively.

### 1. Ten-fold cross-validation test

Because of the wide diversity in which those elements can occur we split the data into 10 parts and performed a 90% learn – 10% test evaluation testing once each part [Kohavi, 1995]. Table 2 shows the results for each of the three element types.

Table 2: Results for extracting addresses, phones and e-mails without sub-elements

	Count	Recall	Precision	Accuracy
Address	134	51.54%	74.56%	57.25%
Phone	296	82.71%	87.45%	69.41%
E-mail	102	89.96%	97.29%	95.44%

During generalization the number of states in the automata has been reduced on average by 71%, 72% and 90% for addresses, phone numbers and e-mails respectively. The automata could be further compacted by merging common sub-trees.

This experiment shows the ability of the learning method to build generalized automata for parsing web documents. There appears to be a relation between the algorithm's performance and the structural variance of the information to be extracted.

### 2. Examination trend of reaching satisfactory results with increasing the cardinality of learning set

In other group of experiments the data set was split in two parts in a random principle – 40 instances were used as a learning set and the rest 60 documents were used as a testing set.

The system was learned using respectively 10, 20, 30 and 40 web documents from the learning set (each set contained the documents of the lower learning sub-set). Each time the testing was provided with all documents from the testing set. The test results were analyzed to obtain values for the numbers of fully extracted, partially extracted and elements that should have been but were not extracted. These experiments were provided in order to examine the trend of reaching satisfactory results. For each case multiple randomized runs have been performed to obtain more stable average values. We assume the address is recognized if its sub-elements, given in the text, are recognized. The telephone number is recognized if the system has recognized at least the phone

code and the number. In several cases the system has recognized the string as a whole without recognizing its sub-elements. Partial recognizing means that some sub-elements are recognized (in particular one of them), but not the element as a whole. For instance only "town", "street", etc.

Table 3 shows the obtained results.

Table 3: Precision for extracting addresses, phones and e-mails when learning sets were 10, 20, 30 and 40 documents respectively

Number of Learning Documents	Address	Phone	E-mail
10	45.33%	85.54%	93.30%
20	50.43%	81.17%	86.72%
30	54.24%	84.23%	88.73%
40	66.19%	85.57%	93.35%

Figures 7, 8 and 9 shows the trend of increasing learning accuracy with increasing of the cardinality of learning set for elements and sub-elements of addresses and phones respectively.

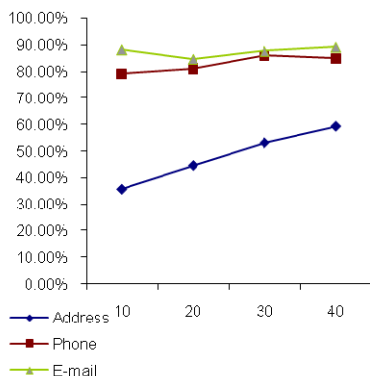


Figure 10. F-measure for addresses, phones and e-mails

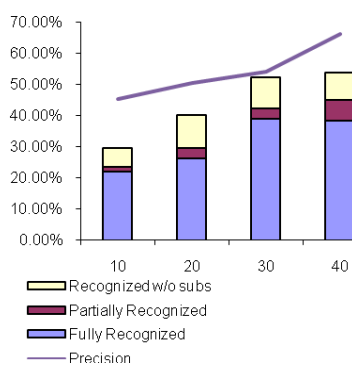


Figure 11. Recall and precision for the addresses

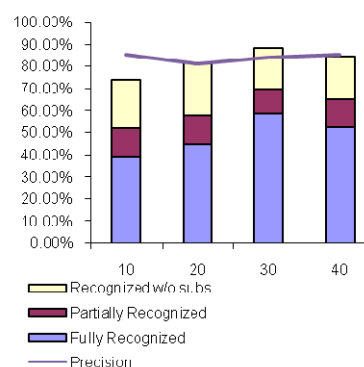


Figure 12. Recall and precision for the phones

In this experiment we found there is a trend for increasing the accuracy of the extractor by increasing the learning data set. As expected, e-mail addresses show the highest recall and precision and achieve high accuracy with a small cardinality of the learning set. The main reason for it is probably the existence of a strict and short structure for an e-mail address which leads to little variety in the different element instances. In that case learning over wider range of documents can actually sometimes prevent the optimal generalization resulting in worse results (Table 3). Bulgarian addresses show the worse results. Given the extremely wide variety of the indirect representation of Bulgarian addresses the results for this element are very promising. Furthermore, by increasing the learning and testing data sets the automata should begin to comprise of the most common cases which will lead to results comparable to the ones for the other two elements.

Our results are compatible with [Cohen, 2004]. The differences are coming from different languages and different grammars' structure in the languages.

---

## Conclusion

---

The aim of the current work was to propose an approach for indirect spatial data extraction by learning restricted finite state automata from marked web documents. The learning method, program realization and experiments were presented. The proposed approach is suitable to cover practical needs for automatic extraction of indirect spatial data from web documents created using Bulgarian language.

The developed system INDES and provided experiments showed that such approach is acceptable and can be used in INSPIRE network.

The main idea of the approach is based on the understanding that the most of indirect spatial information objects are referenced to specific locations or geographic areas using the addresses. In near future, such kind of information will be given following the INSPIRE Data Specification of Addresses [INSPIRE-DSAD, 2008]. It is good standard which is accepted all over the European Community and need to be basis for the further investigation.

The future work involves research in the following directions:

- Adding external knowledge to the system (part-of-speech tagging, named entity lists, word ontology);
- Enhancing the generalization algorithm to identify common sub-trees and merge them if possible;
- more detailed comparison the performance of the realized system with other existing systems like MinorThird which implements various other extractor learning algorithms [Cohen, 2004].

---

## Acknowledgements

---

This work is partially supported by Bulgarian National Science Fund under the project D 002-308 / 19.12.2008 "Automated Metadata Generating for e-Documents Specifications and Standards".

---

## Bibliography

---

- [Baltes, 1992] J. Baltes. Symmetric Version Space Algorithm for Learning Disjunctive String Concepts. Technical Report 92/469/06, University of Calgary, Calgary, Alta, March 1992
- [Ciravegna, 2001] F. Ciravegna. Adaptive Information Extraction from Text by Rule Induction and Generalisation. IJCAI 2001, pp. 1251-1256.
- [Cohen, 2004] W. Cohen. Minorthird: Methods for Identifying Names and Ontological Relations in Text Using Heuristics for Inducing Regularities from Data. 2004. <http://minorthird.sourceforge.net>
- [Holzmann, 1991] J. Holzmann. Design and Validation of Computer Protocols. Prentice Hall, 1991, 512 p.
- [IBM, 2009] Trainable Information Extraction Systems. <http://researchweb.watson.ibm.com/IE/>
- [INSPIRE Directive, 2007] Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)  
<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:EN:PDF>
- [INSPIRE-DSAD, 2008] INSPIRE Data Specification of Addresses  
[http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/DataSpecifications/INSPIRE\\_DataSpecification\\_AD\\_v2.0.pdf](http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/DataSpecifications/INSPIRE_DataSpecification_AD_v2.0.pdf)
- [INSPIRE-DSM, 2007] INSPIRE Data Specifications: Methodology for the Development of Data Specifications  
[http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/inspireDataspecD2\\_6v2.0.pdf](http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/inspireDataspecD2_6v2.0.pdf)

[INSPIRE-TAO, 2007] INSPIRE Technical Architecture Overview.

[http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/network/INSPIRETechnicalArchitectureOverview\\_v1.2.pdf](http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/network/INSPIRETechnicalArchitectureOverview_v1.2.pdf)

[Kohavi, 1995] R. Kohavi. A Study of Cross Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence IJCAI, 1995.

[Li et al, 2008] Y. Li, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, H. Jagadish. Regular Expression Learning for Information Extraction. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, October 2008. Association for Computational Linguistics, pp. 21–30.

[Taylor, 1982] J. Taylor. An Introduction to Error Analysis. University Science Books, Mill Valley, California, 1982.

[Rozenfeld et al, 2008] B. Rozenfeld, R. Feldman. Self-supervised relation extraction from the Web. Knowledge and Information Systems, Volume 17, Issue 1, (October 2008), Springer-Verlag New York, Inc. USA, pp. 17-33.

---

### Authors' Information

---

*Dimitar Blagoev* – Plovdiv University "Paisii Hiledarsk"i, PhD Student in Computer Informatics Department; 24, Tsar Asen St., 4000 Plovdiv, Bulgaria; e-mail: [gefrix@gmail.com](mailto:gefrix@gmail.com)

*George Totkov* – Plovdiv University "Paisii Hiledarski"; chair of Computer Informatics Department; 24, Tsar Asen St., 4000 Plovdiv, Bulgaria; e-mail: [totkov@uni-plovdiv.bg](mailto:totkov@uni-plovdiv.bg)

*Milena Staneva* – Institute of Mathematics and Informatics – BAS; Information Systems Department; Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: [mstaneva@math.bas.bg](mailto:mstaneva@math.bas.bg)

*Krassimira Ivanova* – Institute of Mathematics and Informatics – BAS, Information Systems Department; Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: [kivanova@math.bas.bg](mailto:kivanova@math.bas.bg)

*Krassimir Markov* – Institute of Mathematics and Informatics – BAS, Information Systems Department; Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: [markov@foibg.com](mailto:markov@foibg.com)

*Peter Stanchev* – Kettering University, Flint, MI, 48504, USA / Institute of Mathematics and Informatics – BAS; chair of Information Systems Department; Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: [pstanche@kettering.edu](mailto:pstanche@kettering.edu)

---

## ADAPTATION FOR ASSIMILATION: THE ROLE OF ADAPTABLE M-LEARNING SERVICES IN THE MODERN EDUCATIONAL PARADIGM

Damien Meere, Ivan Ganchev, Stanimir Stojanov, Máirtín O'Dróma

*Abstract:* This paper presents an adaptable InfoStation-based multi-agent system facilitating the mobile eLearning (mLearning) services provision within a University Campus. The network architecture is presented, and the interactions between the various components within the architecture during mLearning service provision (mLecture, mTest) are presented. System implementation approaches are also considered, with particular attention paid to the creation of user profiles and service profiles for the personalization and contextualization of the presented services.

*Keywords:* InfoStations, Intelligent Agents, Multi-agent System, mLearning, CC/PP-UAProf, OWL-S.

---

### I. Introduction

The InfoStation-based system described in this paper is established and operates across a University Campus area mainly for the purposes of the mobile eLearning (mLearning) process. It provides "many-time, many-where" wireless services accessible via mobile devices (cellular phones, laptops, personal digital assistants-PDAs) through geographically intermittent high-speed connections. In this paper we highlight the architecture underlying the provision of these services. We also discuss how the various components of the network architecture collaborate in order to facilitate the mLearning services provision in both blended and distance learning environments. Particular emphasis will be placed on the architectures ability to adapt and customize the services to meet the capabilities of the particular operating environment, as well as personalizing the service to suit a particular end user.

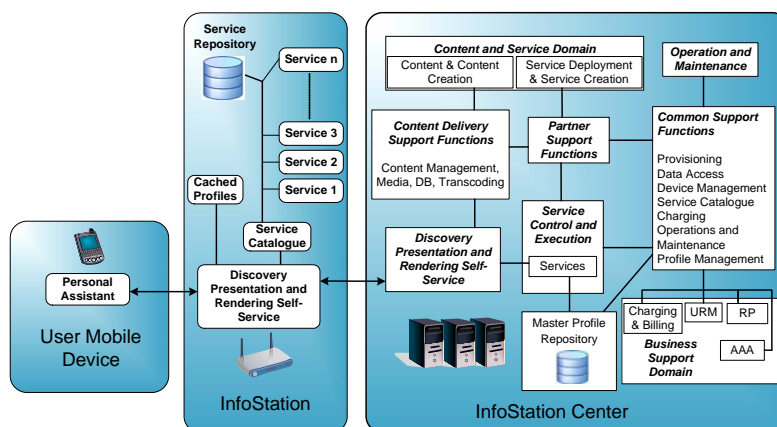
The rest of the paper is organized as follows. Section II presents briefly the InfoStation-based network architecture. Section III illustrates the mLearning service provision outlining sample interactions between system entities. Section IV outlines some implementation issues, in particular regarding the use of semantic web technologies in order to facilitate a personalized and contextualized learning environment through the creation of user profiles and service profiles. Finally Section V concludes the paper.

---

### II. InfoStation-based Network Architecture

The following InfoStation-based network architecture provides access to mLearning services, for users equipped with mobile wireless devices, via a set of InfoStations deployed in key points across the University Campus [1-4]. Whilst within these isolated pockets, within range of an InfoStation, clients may access these localised and personalised mLearning services, in a "many-time, many-where" fashion through a geographically intermittent high speed connection. The InfoStation paradigm is an extension of the wireless Internet as outlined in [5], where mobile clients interact directly with Web service providers (i.e. InfoStations). The evolution in capabilities of, and resources available in modern mobile devices, has precipitated an evolution in the realm of eLearning. The following presented architecture serves to harness this communicative potential in order to present learners with a pervasive learning experience which can be dynamically altered and tailored to suit the learner. Due to the geographically intermittent nature of the connection to the InfoStations, it is necessary for intelligent agents to

operate also onboard the user mobile devices [6-8]. Acting as "Personal Assistants", these agents are able to function autonomously in order to satisfy any user service requests they may encounter, while in or out of contact with other agents (installed on the InfoStations and/or InfoStation Center). This agent autonomy allows the most efficient utilization of the InfoStation's high-rate intermittent coverage. The 3-tier network architecture consists of the following basic building entities as depicted in Figure 1: user mobile devices, InfoStations and an InfoStation Center.



**Figure 1.** The 3-tier InfoStation-based network architecture with some entity components utilized from [9]

The users request mLearning services (from their mobile devices) from the nearest InfoStation via available Bluetooth (IEEE 802.15 WPAN), WiFi (IEEE 802.11 WLAN), or WiMAX (IEEE 802.16) connection. The InfoStation-based system is organized in such a way that if the InfoStation cannot fully satisfy the user request, the request is forwarded to the InfoStation Center, which decides on the most appropriate, quickest and cheapest method of delivering the service to each user according to his/her current individual location, current operating environment (mobile device's capabilities and wireless access constraints) and indeed the preferences of the particular user. The InfoStation Centre maintains an up-to-date repository of all profiles and eContent. The InfoStations themselves maintain cached copies of all recently used user profiles and service profiles, as well as a local repository of cached eContent. In the following section we describe the provision of a number of mLearning services.

### III. mLearning Services

#### mLecture Service

The mLecture service exists as one of the core mLearning services within this system. Initially this service might be used as a supplementary aid to the traditional learning experience. In a traditional learning environment, it is often the case that a lecturer will make available various ancillary learning resources in order to aid the students understand the presented information. Usually this would consist of a set of notes to accompany the lecture. In this case, a lecturer could upload these learning resources (e.g. text, diagrams, images) pertaining to the lecture, or indeed supply a pod-cast or possibly even a video-cast of the lecture itself. This would greatly assist the student's assimilation of information. As students would be able to regulate the pace at which they proceed through the information and if necessary re-cycle back through the lecture. This ensures the material be more

accessible for learners of varying learning styles. Indeed for some time now, pod-casts have been utilized around the world in third level institutions to aid students.

The following, Figure 2, outlines the entity interactions that take place during the provision of the mLecture service. The purpose of this service is to allow students to gain access to lecture content through their mobile devices. The students can request material relevant to specific lectures, which is delivered to their mobile device. However, first this service is adapted and customized according to the capabilities of the user devices and the users own preferences (specified within user profiles). The user device may be limited to the utilization of a text and/or audio only, in which case if there are video components available they will be omitted. The user may choose to access the full capabilities of the service later, whilst using a device of greater capabilities (e.g. PDA, laptop). This trimming (adaptation) of the services is one way to address the shortcomings of some mobile devices, while still delivering the service.

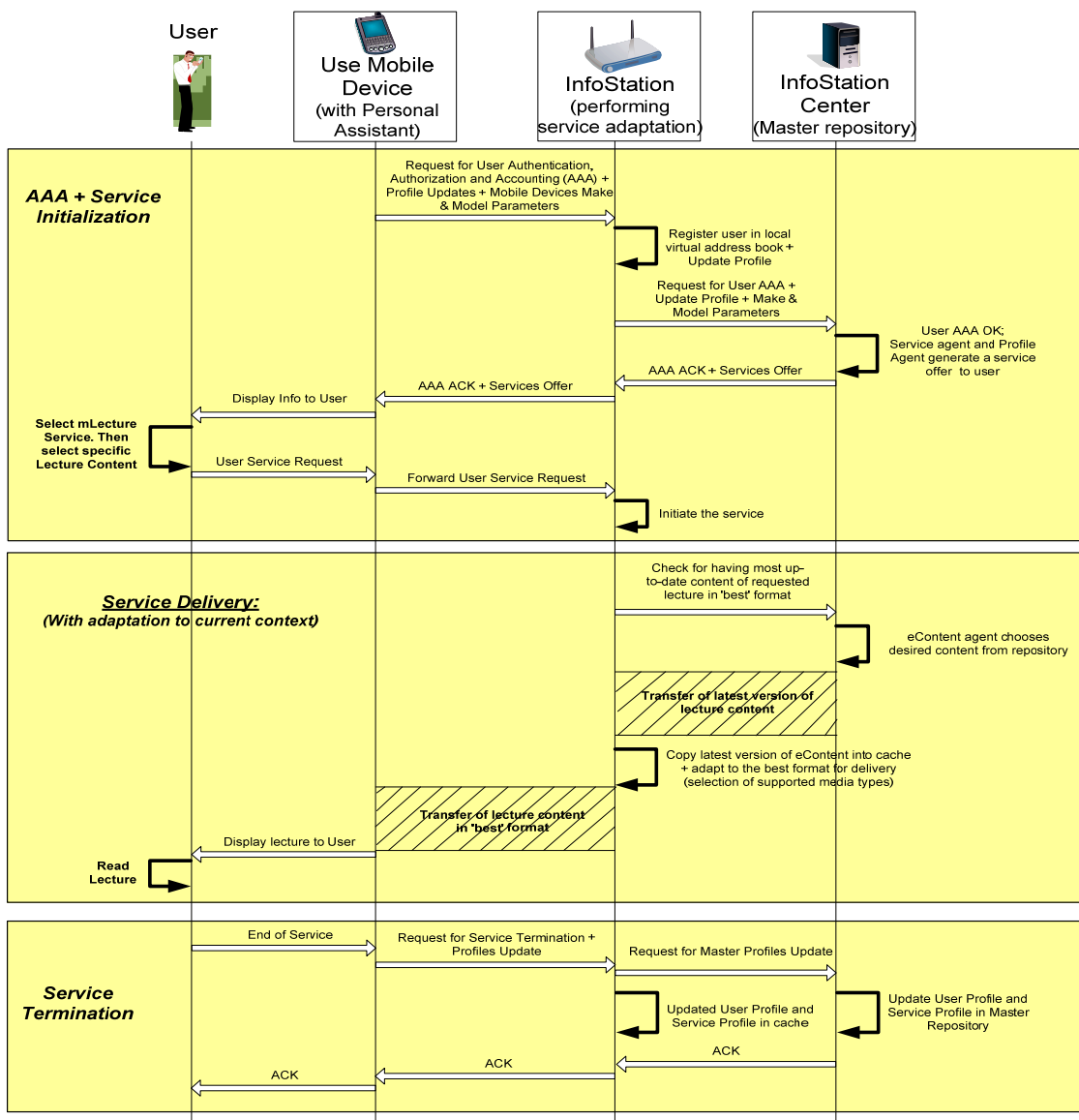


Figure 2. mLecture service provision

---

When the mobile user first enters within the range of an InfoStation, the InfoStation is initially concerned with the processing of the user's AAA request before subsequent selection of service. The InfoStation registers the user's details within its local repository and processes any necessary updates to the user profile and service profile. If an update is necessary, the InfoStation forwards the various profile updates to the InfoStation Center, which can in turn disseminate these updates throughout the InfoStation network. If the Local InfoStation has no record of the user, the user AAA request can be forwarded to the InfoStation Center, which processes the request and facilitates the InfoStation with the requisite profile information. Once the AAA procedure has been successfully completed, the user is presented with a list of available services. Let's assume that the service chosen is the *mLecture*. The PA forwards the user's service request to the InfoStation, which first initiates the services and then checks with the InfoStation Center whether or not it has the most up-to-date cached version of the required *mLecture*. If not, the InfoStation Center sends back a full copy of the requested lecture to the InfoStation. The InfoStation examines the user profile and service profile and customizes the service accordingly. The user's current device may not possess the capabilities to accessing the full range of media formats available as part of a service. For this reason, the InfoStation will adapt the service content to a format which 'best' suits that particular device (i.e. accounting for both the device and the access network) according to the specifications within the User Profile.

#### mTest Service

This service is crucial to the complete eLearning process. mTest provides a means to evaluate the student's acquired knowledge and provides valuable feedback to students concerning their progress. mTest also allows the educator to shape the learning experience of the students, ensuring the student remains engaged in the correct material. Indeed the main benefit of using quizzes is the motivation of the student's engagement in the material, without the stress associated with traditional exams. By providing feedback on their progression, students can be made aware of how well they are assimilating the presented course content. Educators may also benefit from such information. By monitoring the progression of a group of students, the educator may actively modify their approach to conveying the course content and as such, optimize the performance of the group, and enhance the overall learning experience. Nowadays the line between mobile phones and devices such as PDAs and laptops has become so blurred that in some cases there really is little difference in terms of capabilities. As such, the mTest service must be capable of utilizing the full capabilities of the device on which it's being accessed. In addition, more advanced capabilities afford content developers the opportunity to be more creative in designing multimedia mTests. On low-level devices with limited capabilities, a simple text format can be adopted for the creation of the assessments. However on those devices capable of supporting multi-media, assessments may incorporate elements of text, images, sounds and even videos. All of which serve to actively engage students in the material being assessed, especially when utilized alongside the *mLecture* service.

The following sequence diagram, Figure 3, depicts sample interactions between entities involved in this service. As outlined previously, when a mobile user enters within the range of an InfoStation, the InfoStation registers the user's details within its local repository and processes any necessary updates to the user- and service profiles. If an update is necessary, the InfoStation forwards the various profile updates to the InfoStation Center, which can in turn disseminate these updates throughout the InfoStation network. If the Local InfoStation has no record of the user, the user AAA request can be forwarded to the InfoStation Center, which processes the request and facilitates the InfoStation with the requisite profile information. Once the AAA procedure has been successfully completed the user is presented with a list of available services. In this case the *mTest* service is chosen. Once



the user/student gains access to the service, s/he may choose a particular assessment. The PA on the user's mobile device forwards a service request on to the InfoStation, specifying the user's choices. The InfoStation in turn, having analysed the capabilities of the target device, discerns the optimal format in which to present the assessment. This service content is then made available to the PA.

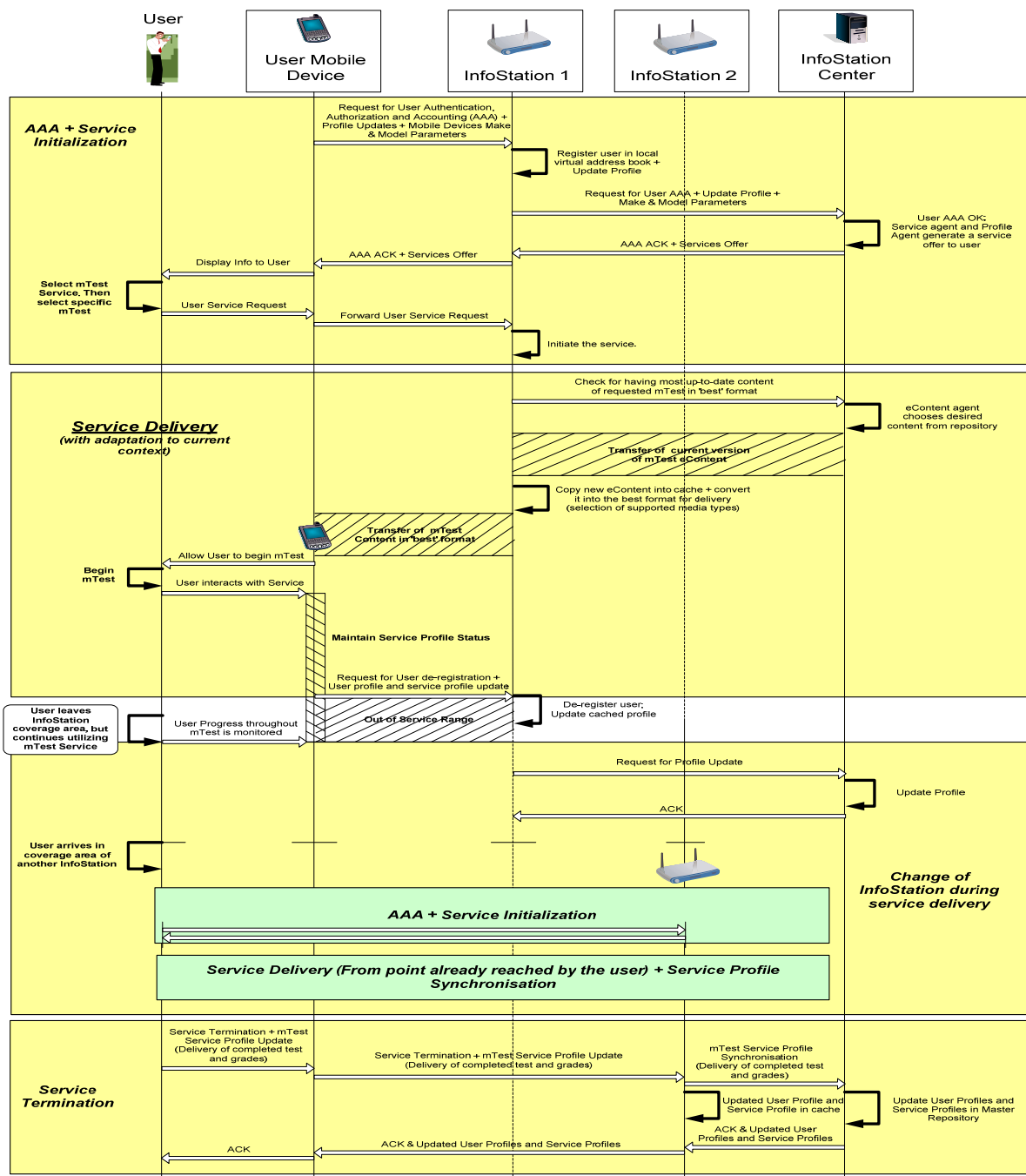


Figure 3. mTest service provision

As the student progresses through the test, his/her user profile is maintained to reflect this progress. Furthermore we consider the possibility for the student to do the test whilst on the move and out of range of an InfoStation. Due to the geographically intermittent nature of an InfoStation connection, the student's mobile device may leave the contact range of the initial InfoStation (InfoStation 1). However the PA facilitates the users continued

utilization of the service while at the same time maintaining the user profile. Thus the student may complete the test whilst outside the radio range of any InfoStation, with the user profile reflecting the student's progression through the material. InfoStation 1, the last InfoStation in contact with the PA, sends user profile updates to the InfoStation Center, ensuring it has the most up-to-date information. Once the PA eventually does enter back within range of an InfoStation (i.e. InfoStation 2), the PA will, after completion of the AAA procedure, forward on a user profile update which reflects the progress of the student through the test content, whilst out of range of the InfoStation network. These updates are disseminated through to the InfoStation Center so as to ensure all information across the system is up-to-date. To the user, this transition between InfoStations should appear seamless, with the user not experiencing any loss of service. Once the user has completed the mTest, the PA displays the results of the assessment to the user, providing valuable feedback on their own progress and performance.

---

#### IV. Implementation Issues

---

##### OWL-S: Service Profiles

We intend to implement the communication between the intelligent agents and the eServices by means of: (i) our own developed communications mechanisms (as explained in the example); and (ii) the standard Ontology Web Language (OWL-S) protocol [10, 11]. The OWL-S is a semantic web protocol that provides a set of constructs with which to create ontologies, which are machine understandable descriptions of the service. These ontologies enable agents to identify and interoperate with the various services. Using OWL-S, the ontology structure is divided into four separate sections, each dealing with a different aspect of the service:

- Service Profile: this advertises the abilities of the service (i.e. what it can do), in such a way as enable a service-seeking agent to determine if the service meets its requirements.
- Service/Process Model: this gives a detailed description of the operation of the service and tells a service user how and when to interact with a service (read/write messages). Essentially it outlines how to initiate the service, and what happens when a service carries out its purpose.
- Grounding: this provides details of how the agent can interoperate with a service (i.e. interact with the service).
- The Service: This simply binds the other elements together into a single entity which can be published and invoked.

When these separate elements are combined, they form an ontology/description that allows intelligent agents to discover, invoke, compose and monitor eServices. Agents utilize the information contained within the *Service Profile* to ascertain whether or not a service meets its requirements, and adheres to certain constraints such as security, and quality etc. The profile serves only to provide a description of the service to a registry. In this system, the InfoStation provides a user's PA with access to a registry of services. Here the user can examine the available service profile information to locate a particular relevant service. Once the user has selected a particular service, the profile has fulfilled its purpose and performs no other function. The *Process Model* provides the information necessary for the agent to use the service. The Process Model allows the agent to perform a more in-depth analysis of the service and its capabilities, and determine if it can be utilized. It informs the PA of how and when to interoperate with the service (read/write messages). The process model is not a program to be executed, but rather a specification of the methods in which a client must interact with a service in order for the desired outcome to be achieved. The Service/Process Model also allows agents to monitor the execution of tasks

---

performed by a service (or a set of services), and to coordinate the entities involved in the service execution. The *Service Grounding* details how agents can communicate with, interoperate with, and invoke a service. Specified within the grounding, are details pertaining to message formatting, transport mechanisms, protocols, addressing etc. When combined with the details outlined within the process model, all the information necessary for a PA to utilize a particular service is presented.

#### CC/PP-UAProf: Device and User Profiles

For the implementation of the User Profile which is integral to the facilitation of fully adaptable services, we have opted to use the uniform format "Composite Capabilities/ Preference Profile" (CC/PP) [12, 13]. This format is platform-independent and is based on the Resource Description Framework (RDF) [14, 15] and is recommended by the World Wide Web Consortium [16]. A CC/PP profile is basically a description of device capabilities as well as specific user preferences that can be utilized to guide the adaptation of service content delivered to that device. This adapted and personalized mLearning allows us to offer multimedia content and activities adapted to learners' specific needs and influenced by their specific preferences and context. So when a specific user / mobile device submits a request to use a certain service, the source of that service (i.e. the InfoStation) customizes and tailors the service content to meet the user preferences and the capabilities of his/her current mobile device. In essence, content is adapted to 'best' suit the individual user and the specific device at that particular time. Through the customization and tailoring of the services (and their content), they can be offered to users, independent of the type of mobile devices. This is an essential factor in this type of network, as user devices and preferences will be as varied as the users themselves. A CC/PP profile contains a number of attributes and associated values, which are used by the InfoStations to determine the most appropriate ('best') format of the resource to be delivered to the user's PA. The User Agent Profile (UAProf) [17-19] specification is a concrete implementation of the CC/PP developed by the Open Mobile Alliance [20]. This specification builds upon WAP 2.0 and facilitates the flow of capability and preference information (CPI) between the Personal Assistant, the InfoStation and the InfoStation Center. This specification defines this capability and preference information through a structured set of components and attributes. The following are the most useful components defined within the UAProf specification. However we could add our own additional components and attributes to better convey capability and preference information within our system:

- *Hardware Platform*: contains attributes that describe the hardware characteristics of the current user device, e.g. device type, input and output methods, screen size, color capabilities, image capabilities, device CPU etc.
- *Software Platform*: contains attributes relating to the operating environment of the device, e.g. operating system name-vendor-version, JVM version, audio & video codecs, Java enabled etc.
- *Network Characteristics*: attributes relating to the network capabilities of the terminal, e.g. bearer characteristics – latency, reliability etc.

The different entities within the system can use this CPI to ensure that the user receives service/content that is tailored for the environment in which it will be accessed. However, it is possible to even further customize the service to suit the preferences of the user. This is achieved through the extension of the CC/PP vocabulary. A CC/PP vocabulary defines the format or structure of the profile information, which is exchanged between a Personal Assistant and the InfoStation. While CC/PP and UAProf already define a number of components and attributes to describe the many different capabilities of the user device, we define a number of attributes relating to the user himself/herself, which could be used to further customize and enhance the service for that individual

user. The user preference components can specify anything from the user's name, the languages s/he speaks, user's age, location, and the format in which the user would prefer to receive information. Another important attribute within the user profile is to specify the role or job title of the user, i.e. whether the individual is an educator or a student etc. Specific groups may be allowed access to different resources related to the service. This is especially within a University environment, where students from different faculties may require access to different services. The following Figure 4 is an example of how a component and group of attributes relating to a particular user may specify vital information about that individual, which can be used to facilitate a higher degree of personalization and quality of service to users.

```

<prf:component>
<rdf:Description rdf:ID="UserPlatform">
<rdf:type rdf:resource="http://www.ece.ul.ie/trc/profiles/UAPROF/ccppschem-1#UserPlatform" />
<prf:Name>John Doc</prf:Name>
<prf:StudentID>0123456</prf:StudentID>
<prf:Faculty>ECE</prf:Faculty>
<prf:Course>Electronic Engineering</prf:Course>
<prf:Year>4</prf:Year>
<prf:Classes>
<rdf:Bag>
<rdf:li>CE4517</rdf:li>
<rdf:li>CE4607</rdf:li>
<rdf:li>CE4717</rdf:li>
<rdf:li>CE4817</rdf:li>
<rdf:li>CE4907</rdf:li>
<rdf:li>EE4607</rdf:li>
</rdf:Bag>
</prf:Classes>
<prf:Advisor>Dr. Ivan Ganchev</prf:Advisor>
<prf:email>0123456@STUDENT.ul.ie</prf:email>
<prf:QCA>3.47</prf:QCA>
<prf:FYP>JD09</prf:FYP>
<prf:FYPSupervisor>Dr. Ivan Ganchev</prf:FYPSupervisor>
<prf:FYPTitle>Design and Implementation of an Animated Interactive Tutorial</prf:FYPTitle>
</rdf:Description>
</prf:component>

```

**Figure 4.** Example of a CC/PP User specific Component

Within this sample profile component, we can see some attributes which will prove essential in order to properly offer services to the most applicable users, and indeed adapt them to suit this particular user. In the given sample component, an attribute such as 'Faculty' can have a major bearing on the type of services being offered to a particular user. It is essential for these factors to be taken into account in order to avoid unnecessarily advertising irrelevant services to users.

## V. Conclusion

The implementation of the adaptable InfoStation-based mLearning service provision within a University Campus has been outlined in this paper. The underlying network architecture has been detailed. The mLecture and mTest services, which provide a means to evaluate the students acquired knowledge and provide valuable feedback to students concerning their progress, has been described. The entity interactions involved in facilitating these services have been detailed. The process of adapting and customizing the service content according to the capabilities of the current user device, current access network constraints and the user preferences has also been outlined. The utilization of the Composite Capabilities/ Preference Profile" (CC/PP) format for the implementation of the User/Service Profiles, which are integral to the adaptation of the services, has been outlined. The benefits of using this format have also been considered.

---

**Bibliography**

---

- [1] I. Ganchev, M. O'Droma, D. Meere, and S. Stojanov, "On Development of InfoStation-based mLearning System Architectures," in *8th IEEE International Conference on Advanced Learning Technologies (IEEE ICALT08)*, Santander, Cantabria, Spain, 2008.
- [2] I. Ganchev, M. O'Droma, D. Meere, and S. Stojanov, "InfoStation-Based Adaptable Provision of m-Learning Services: Main Scenarios," *International Journal "Information Technologies and Knowledge" (IJ ITK)*, vol. 2, pp. 475-482, 2008.
- [3] I. Ganchev, D. Meere, M. O'Droma, S. Stojanov, and M. O'hAodha, "Integrating the Educational Support Architecture in an E-Services Paradigm: The M-Learning Approach," in *M-libraries: libraries on the move to provide virtual access*, G. Needham and M. Ally, Eds.: Facet Publishing, 2008.
- [4] I. Ganchev, S. Stojanov, M. O'Droma, and D. Meere, "An InfoStation-Based Multi-Agent System Supporting Intelligent Mobile Services Across a University Campus," *Journal Of Computers*, vol. 2, May 2007.
- [5] M. Adaçal and A. Bener, "Mobile Web Services: A New Agent-Based Framework," *IEEE Internet Computing*, vol. 10, pp. 58-65.
- [6] N. Sadeh, E. Chan, Y. Shimazaki, and L. Van, "MyCampus: an agent-based environment for context-aware mobile services," in *AAMAS02 Workshop on Ubiquitous Agents on Embedded, Wearable, and Mobile Devices*, Bologna, 2002.
- [7] J. Hendler, "Agents and the Semantic Web," in *IEEE Intelligent Systems '01*, 2001, pp. 30-37.
- [8] FIPA, "Foundation for Intelligent Physical Agents (FIPA) at <http://www.fipa.org>"
- [9] C. Andersson, D. Freeman, I. James, A. Johnston, and S. Ljung, *Mobile Media and Applications - from concept to cash*: Wiley, 2006.
- [10] D. Martin, M. Paolucci, S. McIlraith, M. Burstein, D. McDermott, D. McGuinness, B. Parsia, T. Payne, M. Sabou, M. Solanki, N. Srinivasan, and K. Sycara, "Bringing Semantics to Web Services: The OWL-S Approach," in *1st International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004)*, San Diego, CA, USA, 2004.
- [11] D. Martin, M. Burstein, J. Hobbs, O. Lassila, D. McDermott, S. McIlraith, S. Narayanan, M. Paolucci, B. Parsia, T. Payne, E. Sirin, K. Sycara, and N. Srinivasan, "OWL-S: Semantic Markup for Web Services," W3C 22 Nov. 2004.
- [12] W3C, "Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies 2.0," World Wide Web Consortium (W3C) 8 December 2006.
- [13] L. Tran, M. Butler, E. Izdepski, D. Coward, A. Schade, R. Hermann, S. Chatterjee, and J. Williams, "Composite Capability/Preference Profiles (CC/PP) Processing Specification," Sun Microsystems, Inc October 28 2003.
- [14] W3C, "RDF Primer," 10 February 2004.
- [15] D. Brickley and R. V. Guha, "RDF Vocabulary Description Language 1.0: RDF Schema," World Wide Web Consortium 10 February 2004.
- [16] "World Wide Web Consortium (W3C) at <http://www.w3.org/>" 2009.
- [17] "Wireless Application Group User Agent Profile Specification (WAG UAPROF)," Wireless Application Protocol Forum, Ltd 10 Nov 1999.
- [18] C. Smith and M. Butler, "Validating CC/PP and UAProf Profiles," Hewlett-Packard Laboratories Bristol, UK October 11th 2002.
- [19] "User Agent Profile: Version 2.0," Open Mobile Alliance 6 February 2006.
- [20] "Open Mobile Alliance (OMA) at <http://www.openmobilealliance.org/>" 2009

---

**Authors' Information**

---

*Damien Meere* – Researcher in the Telecommunications Research Centre in the University of Limerick, Ireland. He is currently pursuing his PhD. [Damien.Meere@ul.ie](mailto:Damien.Meere@ul.ie)

*Dr. Ivan Ganchev* – Dip. Eng. (honours), PhD, IEEE (SM.), IEEE ComSoc (SM.); Lecturer and Deputy Director of the Telecommunications Research Centre, University of Limerick, Ireland. He has served on the TPC of many international conferences including IEEE VTC, IEEE Globecom, IEEE ISWCS. [Ivan.Ganchev@ul.ie](mailto:Ivan.Ganchev@ul.ie).

*Dr. Stanimir Stojanov* – Dip. Eng. (Humboldt, Berlin), PhD (Humboldt, Berlin); Associated Professor, Chief of eCommerce Laboratory, and Head of Department of Computer Systems, Faculty of Mathematics and Informatics, University of Plovdiv, Plovdiv, Bulgaria. [S.Stojanov@isy-dc.com](mailto:S.Stojanov@isy-dc.com)

*Dr. Máirtín S. O'Droma* – B.E., PhD, C.Eng., FIEE, IEEE (SM); Senior Lecturer and Director of the Telecommunications Research Centre, University of Limerick, Ireland. He has served on the TPC of many international conferences including IEEE VTC2007Spring, IEEE ISWCS 2006 & 2007. [Mairtin.ODroma@ul.ie](mailto:Mairtin.ODroma@ul.ie)

---

## EULERPATHSOLVER: A NEW APPLICATION FOR FLEURY'S ALGORITHM SIMULATION

Gloria Sánchez-Torrubia, Carmen Torres-Blanc, Leila Navascués-Galante

*Abstract:* EulerPathSolver is a new application, that meets eMathTeacher specifications and simulates Fleury's algorithm execution. The application runs in a Java Web Start Window and features an animation of the algorithm code, a framework working panel showing the algorithm structures and allowing their manipulation, Pop-up questions, language selector and, save/load options together with the interactive simulation within automatic correction of the user's inputs. It has been designed with the main purpose of supporting active learning as well as being a good aid for teachers when explaining the algorithm process. EulerPathSolver enhances dramatically the old Fleury's Algorithm tutorial designed by the authors and will be a great partner when learning Fleury's Algorithm.

*Keywords:* eMathTeacher, eLearning, active learning, interactive Java applications, discrete mathematics learning, algorithm visualization.

*ACM Classification Keywords:* K.3.1 [Computers and Education]: Computer Uses in Education – computer-assisted instruction (CAI), distance learning. K.3.2 [Computers and Education]: Computer and Information Science Education – computer science education, self-assessment. G.2.2 [Discrete Mathematics]: Graph Theory – graph algorithms, path and circuit problems.

---

### Introduction

---

Since about 13 years ago we have been teaching a Discrete Mathematics course (for first semester undergraduate CS students) that includes an extensive graphs chapter. In the meanwhile we have been changing our goals and methodology, turning this subject to mainly algorithm oriented. The background the students bring, both from their previous learning processes and from their social environment, has transformed their reasoning abilities. Our teaching experience shows that Engineering students have significantly increased their algorithmic reasoning capability, both in comprehension and design, tallying with an important decrease of their formal and algebraic reasoning ability. Thus, fostering graphs' algorithmic approach should be decisive on enhancing students' logical potentials.

Moreover, visualization technologies have been proved as a very positive aid to the learning task, when designed and used under the appropriate conditions [Naps 2003a]. However, comprehensive research is required to determine the best methodology to be applied to the design and development of computer-assisted training, as well as the efficiency of the teaching/learning processes based on this particular method of instruction [Hundhausen 2002].

We began by designing a step by step animated visualization tool (simulating Dijkstra's algorithm) [Sánchez-Torrubia 2001], but immediately discovered that our first semester students adopted a passive attitude that was not beneficial at all for their training. Thus we wondered what those tools should accomplish to avoid this passive attitude and concluded that they should act as "nagging" teachers working next to the student. In our opinion, a teacher should encourage learners to get the concept before starting to practice, and, while practicing, should do nothing but expecting the students' response and guide them towards the right solution. This has been our model

---

when defining what a good tool for actively learning mathematics should accomplish. Therefore, our design philosophy, described by eMathTeacher definition and requirements, has always been focused on the student's interactive prediction, i.e. on the algorithm simulation, helped by the visualization.

### eMathTeacher concept

An eLearning tool is eMathTeacher compliant if it works as a virtual maths trainer [Sánchez-Torrubia 2007a]. In other words: it has to be an on-line self-assessment tool that helps students to actively learn mathematic concepts or algorithms by themselves, correcting their mistakes and providing them with clues to find the right solution. The most important feature of these tools is the feasibility to be used in order to practice while the system guides the user towards the right answer. A tool, designed under this philosophy, should also be useful as bLearning (blended learning) complementary material both for being used by teachers in classroom lectures and by students when learning math by themselves.

The aforementioned idea needs to be concreted by means of a list of features that can be implemented. We list the main requirements we have developed to define what a tool should accomplish for attaining this goal.

- *Step by step inquiring*: for every process step, the student should provide the result of the current step execution whilst the application waits in a stand by mode, expecting the user's input.
- *Step by step evaluation*: just after the user's entry, the tool evaluates it, providing a tip in case it is wrong, or executing it when it is ok.
- *Visualization of the step's changes*.
- *Self-assessment levels* option, implementing correction after several steps or even after several iterations.
- *Pop-up questions* assessing the comprehension of the underlying algorithm principles, thus enhancing in-deep understanding and not only process learning.
- *Animated Algorithm code visualization panel*.
- *Framework working panel* showing the current state of the algorithm data structures. It should also allow the user to update those structures.
- *Saving/retrieving* option, data library and/or automatic exercises generation.
- *Language menu*.
- *Easy to use. Clear presentation within a nice and friendly graphic environment*.
- *Flexible and reliable*: allowing users to introduce and modify the example and to repeat the process if desired.
- *Platform independency and continuous availability* (anytime, anywhere).
- *No installation or maintenance* tasks required and low downloading size.

As members of GIDA<sup>2</sup>M, -an education research group- our work is mainly oriented to design and develop applications for actively learning mathematics following the design methodology that we have called eMathTeacher. We have developed several applications, designed under this philosophy, for actively learning graph algorithms. The applications correspond to DFS & BFS, Fleury's, Prim & Kruskal's and Dijkstra's (PathFinder) algorithms [Sánchez-Torrubia 2007a & Sánchez-Torrubia 2009]. Additionally, Mamdani's eMathTeacher [Sánchez-Torrubia 2008b] is a tutorial for actively learning this fuzzy inference method and it has also been designed under the eMathTeacher philosophy. The tools, implemented as Java applications, are available in GIDA<sup>2</sup>M website <http://www.dma.fi.upm.es/gies/gidam/home.html#aplicaciones>



As mentioned above, we have designed several tools, mainly implementing graph algorithms, and have measured their impact on students' learning [Sánchez-Torrubia 2007b]. It has been evaluated by comparing the marks obtained by two different groups of students in the graph exercise of a Discrete Mathematics final exam. In this exercise, the students had to apply two algorithms on a particular graph. The study group used the applications for bLearning while the control group did not use them. Data showed a deeper understanding of algorithms process in the study group, with the highest marks percentage clearly higher.

### Fleury's algorithm and first interactive tool

Fleury's algorithm is designed for finding an Euler Path in an undirected graph. The graph has an Euler path if it is possible to start at a vertex and move along the graph so as to pass along each edge without going over any of them more than once.

In [Euler 1736], Euler states the following result: A finite graph  $G$  contains an Euler path if and only if  $G$  is connected and contains at most two vertices of odd degree. He also sketches a procedure for finding the path consisting on creating a simple circuit, eliminating the used edges, finding new circuits in the remaining graph and joining the new circuits in the proper vertexes. This procedure, very intuitive in theorem's formal demonstration, is not algorithmically effective.

The algorithmic solution to the problem of finding an Euler Path is credited to a French mathematician named Fleury [Fleury 1883]. The basic idea is that when drawing an Euler circuit, all passed edges cannot be used again. So, at any moment in drawing, with all passed edges deleted, the remaining edges must be in one connected component. It starts with a vertex of odd degree —if the graph has none, then start with any vertex—. At each step it moves across an edge which is included in a cycle, unless there is no choice, and then we delete that edge.

The first tool designed under eMathTeacher philosophy simulating Fleury's algorithm (see Figure 1) consisted of an interactive Java applet that implemented it and has been in service for several years. Within this applet the user introduced the graph and simulated the algorithm execution while the application evaluated the provided inputs. In other words: in real time, the applet only evaluated the input introduced by the user. If it was right, the application implemented the order, and then it remained in a stand by mode, waiting for a new one. If the input was not right, an error message appeared on the message window, indicating to the user, what the error was and waiting for the right one. Once the algorithm has been completed, a successful '*end of algorithm*' message was displayed [Sánchez-Torrubia 2008a].

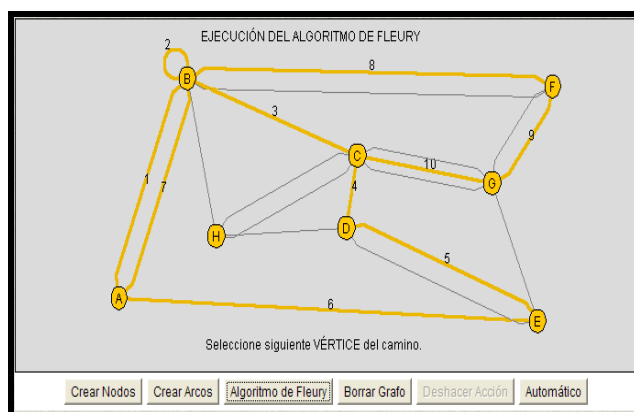


Figure 1: Old Fleury's applet in the process

## New EulerPathSolver application

The applet described in the previous section is a tool that was implemented some years ago and, in the meanwhile, there are some features we lacked. Thus, we have designed a new Java application solving those problems and adding new features:

- One of the Java applets' deficiencies is that, due to security reasons, they are not allowed to write in the client's hard disk. It means that users cannot save their work for later review and every time they want to practice, they have to create the graph. The problem has been solved by using Java Web Start.
- Another feature we missed is related to language capabilities. The first applet was implemented only in Spanish. The new application includes a language menu, starting with Spanish and English and expecting to get some support for being translated to other languages.
- CS students should get used to reading, understanding and executing algorithm code instructions. To achieve this goal with the first tool we encouraged them to read the algorithm code while practicing the algorithm. The Animated Algorithm Code Visualization panel provides a better understanding of the algorithm execution code as it shows the current instruction by changing its colour. Additionally, when the user makes a mistake, the corresponding instruction is shown in red.
- The old tool did not show the algorithm structures at all which made the comprehension of the internal operations of the algorithm difficult. The inclusion of a framework working panel allows users to practice each algorithm step exactly as if they were simulating it by hand, working with the algorithm structures.
- Some of the working details of the algorithm, or the reasons why it works, may remain hidden in the mechanic execution of the pseudo code. The implementation of Pop-up questions, emphasizing these details, will improve the deep understanding of the algorithm.

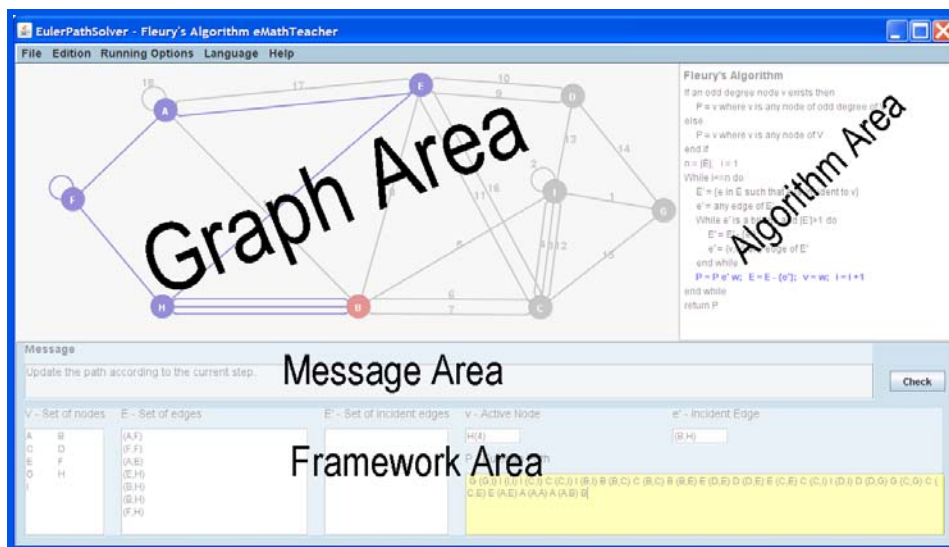


Figure 2: EulerPathSolver new panels

### Tool description

EulerPathSolver is launched by a *jnlp* file and opens in a Java Web Start window.

The window is split into four areas: graph, algorithm, messages and framework areas (see Figure 2).

- The *graph area* displays the graph and allows graph edition. When the algorithm is being simulated, the colours of nodes and edges change according to the algorithm execution.
- The *algorithm area* displays the execution code, showing in blue the current step or in red the point where the user's mistake is located,
- The *message panel* provides clues to find the right solution or indicates the next step to be done. This panel also offers useful hints when the graph is being edited.
- The *framework working panel* shows the structures current state and allows interaction to simulate the algorithm execution.

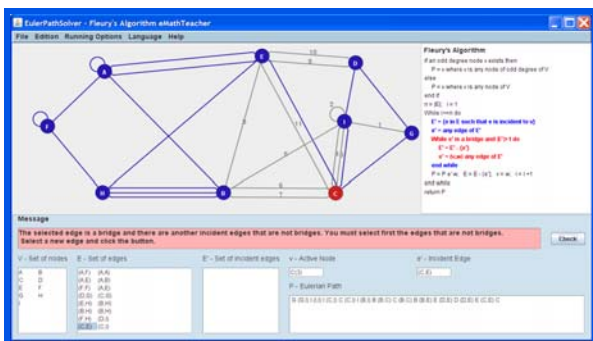


Figure 3: EulerPathSolver showing an error.

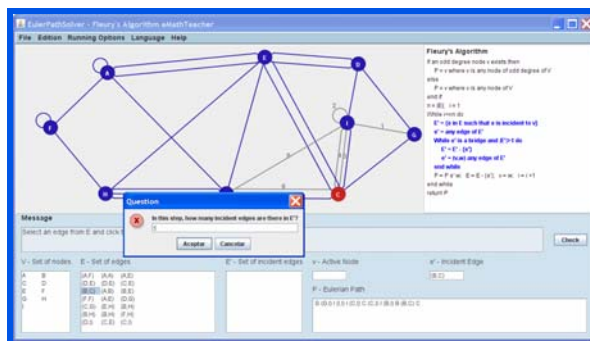


Figure 4: Pop-up question.

The menu bar features a file menu including graph saving/loading and library saving/loading options, edition menu containing create/delete node, create/delete edges, delete graph, undo and redo, two execution modes and a process interruption option. There is also a language selector, currently implemented in Spanish and English.

Once the graph has been introduced and the algorithm is running, the application checks whether an Eulerian path exists. If the graph is not connected or there are more than two odd degree vertexes, an error message is displayed indicating the corresponding fault. If there is an Eulerian path the simulation starts. In each step of the process, the tool shows the current state of the algorithm structures: nodes, edges, active node, edges which are incident with the active node (only in the demonstrative mode) and the covered path.

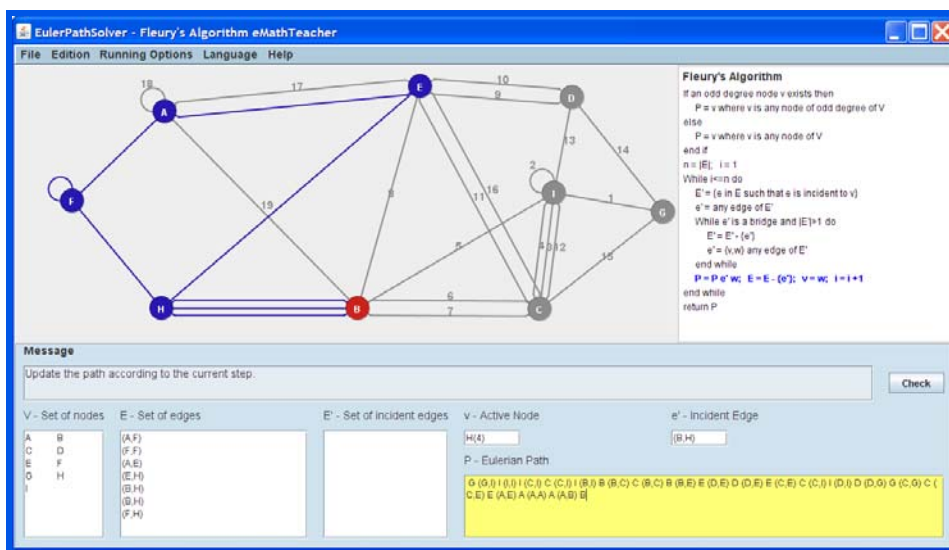


Figure 5: EulerPathSolver updating the Eulerian path.

In the interactive mode, the user should select the next edge to be included in the Eulerian path and, sometimes, add the edge and the new active node to the covered path manually (see Figure 5). For every step, when the user makes a mistake, an error message appears in the message area while the correspondent line within the algorithm code is highlighted in red (see Figure 3). Randomly Pop-up windows appear asking questions related to the current step (see Figure 4).

When the iteration has been completed the application changes the selected edge's color and adds a number indicating the path sequence (see Figure 5).

---

## Conclusion

In the field of graph algorithms we are designing new tools, as well as updating and enhancing the oldest ones. EulerPathSolver meets the specifications listed in the introduction, which means it is eMathTeacher compliant. Actually, the highlighting new feature of this application is an animated algorithm visualization panel, able to display, on the code, the current step and the user's mistake. Other relevant new attribute included is an active framework area for the algorithm data, where the user can modify the algorithm structures while the application verifies the correctness of the input. Finally, it offers the feasibility of saving and retrieving graphs and the graph file follows XML standards.

EulerPathSolver has been designed with the main purpose of supporting active learning (in interactive mode) as well as being a good aid for the teacher when explaining the algorithm process (in demonstrative mode). The new application enhances dramatically the old Fleury's Algorithm tutorial [Sánchez-Torrubia 2008a] and will be a great partner when learning Fleury's Algorithm.

---

## Bibliography

- [Euler 1736] Euler, L. *Solutio problematis ad geometriam situs pertinentis*. *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 8, 128-140 (1736). Based on a talk presented to the Academy on 26 August 1735.
- [Fleury 1883] Fleury. *Deux problemes de geometrie de situation*. *Journal de mathematiques elementaires* 1883, 257-261.
- [Hundhausen 2002] Hundhausen, C. D., Douglas, S. A. and Stasko, J. T. *A Meta-Study of Algorithm Visualization Effectiveness*. *Journal of Visual Languages and Computing*, 13, 3, Elsevier, 259-290 2002.
- [Naps 2003a] Naps, T. L., Rößling, G., Almstrum, V., Dann, W., Fleischer, R., Hundhausen, C., Korhonen, A., Malmi, L., McNally, M., Rodger, S., Velazquez-Iturbide, J. A.: *Exploring the Role of Visualization and Engagement in Computer Science Education*. *Inroads - Paving the Way Towards Excellence in Computing Education*. pp. 131-152, ACM Press, 2003.
- [Sánchez-Torrubia 2009] M.G. Sánchez-Torrubia, C. Torres-Blanc and M.A. López-Martínez, *PathFinder: A Visualization eMathTeacher for Actively Learning Dijkstra's algorithm*. *Electronic Notes in Theoretical Computer Science*, 224 (1), Elsevier, 151-158 (2009).
- [Sánchez-Torrubia 2008a] M.G. Sánchez-Torrubia, C. Torres-Blanc, and V. Giménez-Martínez. *An eMath-Teacher tool for active learning Fleury's algorithm*. *International Journal Information Technologies and Knowledge (IJ ITK)*, 2(5):437-442 (2008).
- [Sánchez-Torrubia 2008b] M.G. Sánchez-Torrubia, C. Torres-Blanc and S. Krishnankutty, *Mamdani's Fuzzy Inference eMathTeacher: a tutorial for Active Learning*. *WSEAS Transactions on Computers* 7 (5), 363-374 (2008).
- [Sánchez-Torrubia 2007a] M. G. Sánchez-Torrubia, C. Torres-Blanc, J. Castellanos, *New Interactive Tools for Graph Algorithms Active Learning*. *ACM SIGCSE Bulletin*, 39 (3), 337 (2007).

---

[Sánchez-Torrubia 2007b] M. G. Sánchez-Torrubia, C. Torres-Blanc, J. Castellanos, Defining eMathTeacher Tools and Comparing them with e&bLearning web based tools. Proceedings of the 2007 International Conference on Engineering and Mathematics (ENMA 2007). (Bilbao, Spain, 7-9 July 2007).

[Sánchez-Torrubia 2001] M.G. Sánchez-Torrubia, V. Lozano-Terrazas, Algoritmo de Dijkstra: Un tutorial interactivo. Proceedings of the VII Jornadas de Enseñanza Universitaria de la Informática (JENUI 2001). (Palma de Mallorca, Spain, 16-18 July, 2001). J. Miró, 254-258.

---

### Authors' Information

---

*Gloria Sánchez-Torrubia* – Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s.n., 28660 Boadilla del Monte, Madrid, Spain; e-mail: [gsanchez@fi.upm.es](mailto:gsanchez@fi.upm.es)

*Carmen Torres-Blanc* – Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s.n., 28660 Boadilla del Monte, Madrid, Spain; e-mail: [ctorres@fi.upm.es](mailto:ctorres@fi.upm.es)

*Leila Navascués-Galante* – Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s.n., 28660 Boadilla del Monte, Madrid, Spain; e-mail: [leila.navascues@gmail.com](mailto:leila.navascues@gmail.com)

## AUTOMATIC METADATA GENERATION FOR SPECIFICATION OF E-DOCUMENTS – THE METASPEED PROJECT

Juliana Peneva, George Totkov, Peter Stanchev, Elena Shoikova

*Abstract: A Bulgarian research project funded by the Bulgarian National Science Fund under the thematic priority Information and Communication Technologies, contract D002-308/19.12.2008 is presented in this article. The main goal and tasks are outlined. Some already achieved results are pointed.*

---

### Introduction

---

In today competitive business environment the proper management of organizational digital resources is crucial for making timely decisions and responding to changing business conditions. Many companies are realizing a business advantage by managing successfully their business data. Resources include documents, images, video or audio clips, animations, presentations, online courses, web pages, etc. Organizations are of different types and sizes ranging from SME to international corporations. All of them exhibit an intensive use of digital resources because these e-documents are stored, distributed, shared and reused without difficulty. Certainly some barriers like technical incompatibility or missing files are to be overcome to achieve an effective use. However digital resources are increasingly being recognized as a very important organizational asset au par with finance and human resources.

In order to be easily retrieved, shared and used from different users and for different purposes the various types of e-documents have to be described following common schemas and rules e.g. specifications/standards and metadata. Depending on content and context standards for e-learning (SCORM, IMS, LOM, etc.), for multimedia data (MPEG-7), to name a few, have been proposed. As a rule, every standard requires too much metadata. Standardized metadata enables the easy choice of relevant e-documents. However poor quality or non-existent metadata means that resources remain invisible within a repository or archive thus becoming undiscovered and inaccessible. On the other hand quality metadata can be produced by experts in the subject domain only. So, building digital repositories with "standardized" e-documents appears to be a labor-consuming, highly qualified and expensive activity. With digital resources being produced in ever-increasing quantities, finding the time and resources necessary for ensuring quality metadata becomes a challenging task. Automation seems promising to address this. We are convinced that automatic metadata extraction could be a right solution. Several approaches, including metatag harvesting, content extraction, automatic indexing or classification, text and data mining, etc. have been proposed. In [1] the quality of currently available metadata generation tools has been compared. The best scenario would be to auto-generate high-quality resource discovery metadata without any human intervention. Nevertheless most of the resource discovery metadata is still created and corrected manually either by authors, depositors and/or repository administrators.

At the same time we would like to mention that in Bulgaria there are no standards or even commonly accepted specifications to describe metadata for e-documents<sup>1</sup>. Again there exist an increasing number of newly created digital repositories in different subject areas e.g.:

---

<sup>1</sup>As a first step in this direction consider the recently adopted Bulgarian Law on e-documents.

- 
- cultural heritage [2] – collections, museum exhibits, old-printed books, science publications, etc.;
  - education [3] – e-learning courses, multimedia learning content, tests items, etc.;
  - spatial information systems [4].

Besides the possibility of applying some well-known world standards such as SCORM, IMS, MPEG-7, it is not expected that the shared e-documents will be specified in a uniform way. This justifies our research efforts namely to automate the process of metadata generation for different in style e-documents. Taking into account the rapidly growing number of new digital repositories investigations in this area are promising.

---

### What is the METASPEED Project?

---

Metadata ExTraction for Automatic SPEcifications of E-Documents – METASPEED is a Bulgarian research project funded by the Bulgarian National Science Fund under the thematic priority: Information and Communication Technologies. It aims to facilitate the development of Bulgarian standards and even commonly accepted specifications for the description of metadata for e-documents in different subject areas. Project partners include Bulgarian researchers from state and private universities and Bulgarian Academy of Sciences. This project is carried out by a consortium composed of: University of Plovdiv, Institute of Mathematics and Informatics – Bulgarian Academy of Science, Technical University of Sofia and New Bulgarian University.

The goal of this project can be briefly summarized as follows: to investigate and create technologies, methods and tools for automatic generation of metadata thus facilitating the proper specification of documents with different e-format, content and location. The rationale behind this goal is that e-documents are to be described in a standardized manner to facilitate their retrieval, sharing and using. Usually documents in digital repositories are determined according to a particular specification and/or standard together with data about the document itself, i.e. metadata. As a rule the application of any standard requires too much metadata that are produced by experts in the subject area. Consider the electronic resources e.g. tests, learning content, etc. in the National Educational Portal [5]. These resources obey no unique standard. That is why our research efforts towards a standardization and automatic generation of metadata for different format e-documents are an economically motivated activity. Project findings will facilitate the access to different digital collections in a straightforward manner. This is the first stage toward the development of a uniform information environment in Bulgaria. We expect that the main contributions include:

- development of proper tools for an automatic metadata generation for collections containing digital documents of different shapes and types;
- building a framework to share European and Bulgarian e-resources;
- development of national standards for document sharing.





---

### Partners

---

The METASPEED project is an interdisciplinary project. This justifies the participations of people interested in computer linguistics, e-learning, standards for e-documents, multimedia applications, archival sciences, database systems, etc. The partners and their competences are presented in Table 1:

Table 1 Partners of the METASPEED Project

	PARTNER	COMPETENCES
	University of Plovdiv Department of Computer Informatics	computational linguistics, standards and systems for e-learning, theory of algorithms, programming languages
	Institute of Mathematics and Informatics - BAS Department of Information Systems	analysis, synthesis and retrieval of structured data from texts, images and video
	New Bulgarian University Department of Informatics	cognitive sciences, database systems, e-learning
	Technical University of Sofia Research laboratory "Technologies and standards for e-learning"	standards and systems for e-learning

---

### Project Work Packages

---

The project is built up of four work packages. Certainly significant dissemination and supporting activities are foreseen.

---

#### *WP1. Standards of e-documents and tools for their automatic generation*

---

The goal of this package is to finalize the research analysis in the area and to prepare state-of the-art reports concerning:

- a. standards for e-learning;
- b. standards for multimedia documents in the field of cultural heritage;
- c. prescriptions for Bulgarian standards in different subject areas;
- d. tools for automatic metadata generation.

It is expected that project technical prescriptions of national standards for e-learning and specifications for cultural heritage e-objects will be proposed. For example the adoption or development of a standard, in the field of e-learning, could provide sharing (including export/import), multiplication and adaptation of learning resources (courses, materials and tests) hosted in the Internet. During the adaptation of well-known standards and specifications a reasonable question arises: to what extent the national specifics such as language, educational system and traditions are to be considered.

---

#### *WP2. Automated metadata generation from text documents*

---

The goal of WP2 is to develop methods, algorithms and tools to retrieve structured data from electronic text documents, written in different languages taking into account existing standards and specifications. To realize this goal a review-analysis of the existing methods and algorithms in the world and in particular – for Bulgarian language will be carried out.



---

The following tasks will be performed:

- a. critical analyses of existing methods and tools for retrieval of metadata from electronic texts;
- b. investigations on specialized technologies and methods for metadata retrieval for documents in different areas;
- c. design and implementation of proper software prototypes;
- d. experiments with the realized methods and tools for metadata retrieval on the documents in examined areas.

---

### *WP3. Automatic metadata generation from multimedia documents*

---

In the next three years, the world will create more data than has been produced in all of human history. It is well known that the search power of current searching engines is typically limited to text and its similarity, since less than 1% of the Web data is in textual form, the rest being of multimedia/streaming nature, particularly since a large portion of pictures still remains as "unstructured data". The Enterprise Strategy Group [6] estimates that more than 80 milliard photographs are taken each year. Using of digital images promises to emerge as a major issue in many areas, for instance Google answers daily more than 200 million queries against over 30 milliard items. Because of this it is necessary to extend next-generation search to accommodate these heterogeneous media.

Some of the current engines search the data types according to textual information or other attributes associated with the files. An orthogonal approach is the Content-based Image Retrieval (CBIR). It is not a new area – in current surveys can be counted more than 300 systems, most of them exemplified by prototype implementations. The typical database size is in the order of thousands of images - very recent publicly-available systems, such as ImBrowse [7], Tiltomo [8] and Alipr [9] declare to index hundreds of thousands of images.

The user questions in image search are partitioned into three main levels:

- a. *Low level* – this level includes basic perceptual features of visual content (dominant colors, color distribution, texture pattern, etc.);
- b. *Intermediate level* – this level forms next step of extraction from visual content, connected with emotional perceiving of the images, which usually is difficult to express in rational and textual terms. Visual art is an example, where these features play a significant role. Typical features in this level are color contrasts, because one of the goals of the painting is to produce specific psychological effects in the observer, which are achieved with different arrangements of colors;
- c. *High level* – this level includes queries according to rational criteria. In many cases the image itself does not contain information which would be sufficient to extract some of the characteristics. For this reason current high-level semantic systems still use huge amount of manual annotation.

Usually the existing systems for image retrieval are limited by the fact that they can operate only at the primitive feature level, while users operate at a higher semantic level. This mismatch is often referred as a semantic gap. The Project aims at finding and analyzing new content-based image retrieval methods to analyze, index, and retrieve images and video. The goal is to increase the retrieval effectiveness by a proper choice of image features from the MPEG-7 standard and on that base to find the description of some concept, which humans use in their everyday life.

---

*WP4. Automated creation and testing of digital repositories in different areas: A) cultural heritage; B) e-learning; C) spatial information systems; D) automated referring of scientific publications.*

---

The main task is to investigate and create methods and tools for automated metadata generation from Web pages (in the listed above subject areas). Known technologies for search in Internet-pages, using tools for automated metadata generation from text and multimedia will be examined and adapted. As a result the Internet-space will be used as a source for building digital repositories in order to test the proposed methods and tools. In addition some collections of electronic resources being created from the Project partners for carrying out experiments in the field of e-learning and scientific publications will also be used to check the developed tools for automated metadata generation.

As it concerns cultural heritage the focus of investigations will be on search methods to be applied in collections of documents containing text and graphical objects. It is expected these documents to be automatically classified following proper ontology.

Investigations in e-learning will deal with automated metadata extraction from learning resources. The possibility for generation of new learning objects taking into account different learning styles will be examined closely. Besides a linguistic analysis of the created learning content, the interactions among trainees and "instructor-trainee" will be studied. The latter facilitates data extraction necessary to build the trainee's portfolio.

The INSPIRE standards for spatial meta-data are obligatory for the European Union member states. This means that it is very important the cultural space information to be described following these standards. Two main tasks are connected to this problem:

- to create Bulgarian standards and thesauri for spatial meta-data of the cultural heritage objects, which are corresponding to the INSPIRE standards;
- following these standards, to develop methods and tools for meta-data extraction from the cultural objects' descriptions.

As a result it is expected that a digital repository of scientific publications will be built via automated metadata retrieval from full-text articles. Referring is a time consuming task because experts have to review many papers and to classify them properly. Again automation seems to be appropriate. Automated metadata retrieval from the collection of scientific publications in different languages (in the case of specialized texts in concrete scientific domain) and their classification according to proper ontology is important in the case of multilingual referring journals (for instance – containing reviews in English, Russian and Bulgarian languages) as well.

Several dissemination and valorisation activities will also be carried out within the frame of the Project. The results from the PhD studies as well as the applied methods will be discussed during some Project specific workshops. It is supposed that most of the results will be presented and approved during the planned project meetings together with our scientific consultants. These results will also be presented as papers on different international and local conferences. When the Project is put into effect, it is presumed that not only extra young scientists will join these prospective investigations but these young people will receive moral and financial incentives to realize their work. At the same time the PhD students will gain the advantage to be in touch with leading researchers in order to acquire their experience. It is possible to provide students' mobility for a short period of time on the basis of a preliminary signed agreement. The organization of specializations for the Project staff and for university lecturers on the standardization in the area of e-learning and metadata generation is one of the aims of the Project management. At least five PhD theses will be promoted within the frame of this Project. In addition students enrolled in masters programs from Plovdiv University, Technical University of Sofia and New Bulgarian University will join the investigations. We expect to promote more than 20 master theses in the Project topics.

---

## Project Deliverables and Overall Effect

---

The proposed in the Project scientific investigations are performed for first time in our country. Some of the proposed approaches can be also considered as an innovation as they have no analogue abroad. The qualification of the participants in the project team and their scientific achievements (including the successful participation in over 60 national and international projects) represent real preconditions for a successful realization of the Project. Last but not least it is expected that the research being carried out in different thematic areas will be very effective. So, the main goals of this Project can be definitely reached.

The project deliverables result from the application of intellectual technologies, in the making of proper models and methods followed by their testing and validation. The software prototypes that would be developed within the Project can be applied not just to test the theoretical results or to prove the models relevance. These prototypes allow for an objective analysis and further improvement of the proposed solutions. The Project deliverables can be summarized as follows: new methods and information technologies tools to be used both in theory and practice in four areas namely e-learning, cultural and historic heritage, spatial information systems, and scientific reviewing, will be worked out.

The proposed problems for investigation are innovative and their solution represents not only matter of a scientific interest. The findings will be used in practice to develop proper methods and tools for an automated metadata extraction and generation from different kinds of e-documents.

It is expected that one of the most important result from this Project concerns the making of technology and methods for automated maintenance of digital repositories (see WP2 and WP3) and all corresponding activities such as: generation and update of virtual catalogues; organization of content-based search; removal of discrepancies, etc. The developed methodology and tools will be tested in the project thematic areas mentioned above (WP4). The success of the Project investigations would overturn the traditional view on the ways of how to plan, organize and maintain a digital archive. Taking into account the Project results the efficiency in the development of new multimedia repositories as well as in updating existing collections of e-documents will be considerable increased.

It is worthwhile to notice the opened possibilities for a wide multiplication of the results. This is especially true as it concerns the application of the developed methods and proposed specifications to generate metadata of different types (content or/and context-of-use based). As an immediate consequence a significant improvement of quality for newly created or hared in different areas digital repositories, will be achieved. Two important deliverables from this Project could be mentioned: an easy-to-use model of a multimedia archive that enables metadata generation, and a research methodology to catalogue digital archives.

Researchers and lecturers from humanitarian and social spheres will be introduced to the problems and the obtained results. The dissemination will be done via specialized seminars and lectures and through Project members' participation in different kind of scientific events.

There are some objective risk factors for the successful finalizing of the Project and the fulfillment of the planned tasks:

- Differences in the organization and the specificity of the research, performed by the partner institutions;
- Insufficient national experience in creating systems and standards for e-documents and, as a consequence, a necessity for exploring and adapting good European and worldwide practices;
- Orientation of the research following in advance fixed goals instead of conforming to its internal logic and some result-driven investigations.

- Involvement of a comparatively large number of young scientists, without proper experience (and enough motivation) to perform long term scientific investigations, etc.

Overcoming the listed above risk factors appears impossible within the frame (and with the resources) of the Project only. However some Project activities included in different work packages e.g. wide publicity; motivation (moral and material) of young scientists and their affiliation to long-term scientific research; the involvement of affiliated scientific consultants with their experience and knowledge; regular (half-year) workshops; open seminars (delivered by experts and representatives of outer organizations), etc. could reduce the influence of the risk factors significantly.

In order to management the Project effectively and to evaluate the progress in every work package a fixed number of milestones are set. They favor a successful finalizing of the corresponding tasks and activities. Examples of milestones include surveys, specifications for a national standard, tools for an automated generation of metadata, papers, PhD theses, a Web portal, presentations, etc.

---

### Dissemination of Knowledge and Expected Impacts

---

The planned and expected project results comprise innovation technologies and methods for generation of metadata. Among them prototypes of software tools that can facilitate the automatic building up of virtual repositories for digital documents, document sharing, repositories maintenance and management are of special importance. It is expected that similar tools possessing peculiar features will appear on the national and European markets for first time. Considering the Project findings a methodology how to build an integrated information repository for digital documents in Bulgaria will be set up. This methodology will be experimented on different thematic areas (see WP4). The main methodology components follow a systematic approach that allows for:

- development and introduction to proper national standards (see WP1);
- shared use and development of digital repositories for e-documents that conform to the proposed standardized specifications (see WP4);
- development of Digital Repository Management Systems to deliver functionalities instead of human experts (see WP3 and WP4), etc.

When the main goals of this Project will be achieved and the planned results become real, further work consist of building an integrated national informational environment to become part of the overall European one.

The Project deliverables (methods, tools and other findings) could serve a sound foundation and proper framework to develop national standards for e-learning and storage of digitalized objects belonging to the cultural and historical heritage. Moreover it becomes feasible to design and build a national informational network to share the repositories' digital content among different institutions e.g. universities, libraries, museums, public archives, community centers, etc.

In consequences of the project results, the efficiency in set up and maintenance of digital repositories will increase because of the possibility for an automated metadata generation and metadata replacement. The latter implies a new research problem: following the proposed methodology the existing standards being used for e-documents description are to be investigated. In addition the developed tools for metadata generation are to be tailored to the new thematic area (see WP2 and WP3).

After finalizing up the Project the team will make efforts to disseminate and valorize the project deliverables. Enterprises and institutions that attend to e-government, developers of digital repositories in various subject areas, civil organizations of the information society, to name a few, could be considered as prospective users of

---

the developed tools and technologies. It has to be mentioned that inherent conservatism in education, information services, etc. gives rise to possible problems in carrying out these valorization activities.

The benefit influence of our research activities on the quality of education in humanities becomes obvious. In humanities new subjects concerning the standardization, storage and digitalization of the object being studied, their representation and exploration via Internet can be introduced. A new possibility for classification and virtual representation of their artifacts via the developed within the frame of the Project tools becomes feasible.

---

## Conclusion

---

In this paper we report about a Bulgarian research project: Metadata ExTraction for Automatic SPEcifications of E-Documents – briefly the METASPEED project, funded by the Bulgarian National Science Fund under the thematic priority: Information and Communication Technologies, contract D002-308/19.12.2008. The main project goal as well as the basic working packages, the overall effect and the expected impacts have been summarized.

Nevertheless that the project is in its very early stage some results have been achieved. Experimental tools and software prototypes to be applied for an automated metadata extraction from text and multimedia content have been developed. Some initial program tools, which will be used in the process of automated metadata extraction from text and multimedia content, are already made.

A system for extracting data from web documents based on simplified finite state automata is developed [10]. Using it custom data structures to be extracted from the documents are specified. The system 'learns' the automata from examples in the form of annotated texts and uses heuristics to expand and improve the automata-extractor. The system is expected to be suitable for the extraction of structured data and metadata in particular.

An environment for modeling and running multilevel processes, which can be used in further analysis of spatial and e-learning metadata extraction, based on expanded variants of Petri-net theory, is proposed. The system is found to be useful for upgrading existing learning management systems by introducing graphical modeling of the e-learning activities workflow [11]. The relevant standardization activities and initiatives, aiming to support the European-wide interoperability of e-learning systems are presented [12].

The main metadata standards of files containing digital photo images are discussed in [13]. Software system realizing extraction of metadata about image content and context is presented. The system gives the opportunity for searching photo images with different criteria using metadata standards EXIF, IPTC and XMP Possible applications of the system are intelligent Internet searching of digital photo images, automated filling in of SCORM metadata (if the corresponded learning material is a digital image), etc.

A classification machine learning system using multidimensional numbered information spaces is built [14]. Some practical implementation of MPEG-7 descriptors for definition and automatic extraction of color harmonies and contrasts, which cover intermediate level of image search are investigated [15]. The work over the area of classification system construction and experiments with trying for automated recognition of high level metadata, based on content based image retrieval, born some questions, which solving led to creating of specific theory that connects categorization/metadata and logic-combinatorial structuring/clustering of the descriptive part of the input table [16]. Some algorithms for group decision making has also been reported [17].

---

## Acknowledgements

---

This work is partially granted by Bulgarian National Science Fund, Ministry of Education and Sciences in the frame of the project "Automated Metadata Extraction for e-documents Specifications and Standards", contract No: D002(TK)-308/ 19.12.2008.

---

## Bibliography

---

- [1] Polfreman M., Rajbhandari S. MetaTools - Investigating Metadata Generation Tools. JISC Final report, October 2008
- [2] Dobreva M., Ikonov N. The Role of Metadata in the Longevity of Cultural Heritage Resources. In Proc. of EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Greece, 2009.
- [3] Larran Aga. et al. Building Learning Objects from Electronic Documents. In Proc of the 16 International Conference on Computers in Education ICCE 2008, Taiwan, October 2008, pp.141-146.
- [4] INSPIRE'07, Directive 2007/2/EC of the European Parliament. Official Journal of the European Union, 25.4.2007, L 108/1. [http://www.epsplus.net/content/download/3477/38314/ile/\\_10820070425en00010014.pdf](http://www.epsplus.net/content/download/3477/38314/ile/_10820070425en00010014.pdf).
- [5] <http://start.e-edu.bg/>
- [6] <http://www.enterprisestrategygroup.com/management>
- [7] <http://media-vibrance.itn.liu.se/>
- [8] <http://www.tiltomo.com/>
- [9] <http://www.alipr.com/>
- [10] Blagoev D., G. Totkov, Information Extraction by Learning Restricted Finite State Automata from Marked Web Documents, iTech'09 (this conference).
- [11] Indzhov Hr., D. Blagoev, G. Totkov, *Executable Petri Nets: Towards Modelling and Management of e-Learning Processes*, ACM International Conference Proceeding Series; Vol. 375, Proc. of the 10th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing 2009, Rousse, Bulgaria, June 18-19, 2009 (in print).
- [12] Doneva R., G. Totkov, N. Kasakliev, E. Somova, European Standardization: Mobility without Borders, ACM International Conference Proceeding Series; Vol. 375, Proc. of the 10th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing 2009, Rousse, Bulgaria, June 18-19, 2009. (in print).
- [13] Totkov G., E. Somova, Hr. Petrov, About Relationship between Metadata and Content of Digital Photo Images, 7<sup>th</sup> International Conference on Emerging e-Learning Technologies and Applications, Nov. 17-20, 2009, Stara Lesna, The High Tatras, Slovakia (accepted).
- [14] Mitov I., Ivanova Kr., Markov Kr., Velychko V., Vanhoof K., and Stanchev P.. PaGaNe – An Intelligent System Based on the Multidimensional Numbered Information Spaces. Fourth Int. Conf. on Intelligent Systems and Knowledge Engineering, 27-28.11.2009, Hasselt, Belgium (to appear).
- [15] Ivanova Kr. and Stanchev P. Color Harmonies and Contrasts Search in Art Image Collections, First Int. Conf. on Advances in Multimedia MMEDIA 2009, 20-25.07.2009, Colmar, France (to appear).
- [16] Ivanova Kr., Mitov I., Markov Kr., Stanchev P., Vanhoof K., Aslanyan L., and Sahakyan H. Metric Categorization Relations Based on Support System Analysis. Seventh Int. Conf. on Computer Science and Information Technologies CSIT 2009, 28.09-02.10. 2009, Yerevan, Armenia (to appear).
- [17] Andonov F. Interactive Methods for Group Decision Making. In Int. Book Series "Information Science & Computing" – Book No: 10. Intelligent Support of Decision Making. Sofia, 2009, pp. 25-30

---

## Authors' Information

---

*Juliana Peneva* – New Bulgarian University, Department of Informatics; e-mail: [july\\_peneva@abv.bg](mailto:july_peneva@abv.bg)

*George Totkov* – University of Plovdiv; chair of Computer Informatics Department; e-mail: [totkov@uni-plovdiv.bg](mailto:totkov@uni-plovdiv.bg)

*Peter Stanchev* – Kettering University, Flint, MI, 48504, USA / Institute of Mathematics and Informatics – BAS; chair of Information Systems Department; Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: [pstanche@kettering.edu](mailto:pstanche@kettering.edu)

*Elena Shoikova* – Technical University of Sofia; chair of Research laboratory "Technologies and standards for learning"; e-mail: [shoikova@tu-sofia.bg](mailto:shoikova@tu-sofia.bg)

## ITA 2010: JOINT INTERNATIONAL SCIENTIFIC EVENTS ON INFORMATICS

June 21 - July 03, 2010, Varna (Bulgaria)

### ITA 2010 Main Scientific Events:



KDS

XVI<sup>th</sup> International Conference "Knowledge – Dialogue – Solution" is traditionally dedicated to discussions on current research and applications regarding three basic directions of intelligent systems development: Knowledge Processing, Natural Language Interface, and Decision Making.



i.Tech

Eight International Conference "Information Research and Applications" is devoted to discussion of current research and applications regarding the basic directions of computer science.



MeL

Fifth International Conference "Modern (e-) Learning" is oriented to discussion of current research and applications in the basic directions of modern (e-) learning: Philosophy and Methodology of the Modern (e-) Learning and Modern (e-) Learning Technologies.



INFOS

Third International Conference on Intelligent Information and Engineering Systems is devoted to current research and applications regarding the knowledge-based intelligent information and engineering systems.



CFDM

Second International Conference "Classification, Forecasting, Data Mining" is directed to the fields of Classification, Clustering, Pattern Recognition, Forecasting, Features Processing, Transformations, Data Mining, and Knowledge Discovery.



ISSI

Fourth International Summer School on Informatics is devoted to discussion of current research, applications and education regarding the basic directions of informatics.

### ITA 2010 is organized by:

ITHEA International Scientific Society

Association of Developers and Users of Intelligent Systems (Ukraine)

Association for Development of the Information Society (Bulgaria)

Institute of Mathematics and Informatics, BAS (Bulgaria)

Institute of Information Technologies, BAS (Bulgaria)

National Laboratory of Computer Virology, BAS (Bulgaria)

Institute of Cybernetics "V.M.Glushkov", NAS (Ukraine)

Institute of Mathematics of SD RAN (Russia)

Institute of Information Theories and Applications FOI ITHEA (Bulgaria)

Dorodnicyn Computing Centre of the Russian Academy of Sciences

Astrakhan State Technical University (Russia)

Ben Gurion University (Israel)

Kharkiv National University of Radioelectronics, (Ukraine)

National University of Kiev "Taras Shevchenko"(Ukraine)

Rzeszow University of Technology (Poland)

Universidad Politecnica de Madrid (Spain)  
University of Calgary (Canada)  
University of Hasselt (Belgium)  
Varna Free University "Chernorizets Hrabar" (Bulgaria)

ITA 2010 is supported by:



The International Journal "Information Theories and Applications"



The International Journal "Information Technologies and Knowledge"



The International Book Series "Information Science and Computing"

### Deadlines

- March 31, 2010: submission of final paper via submission system <http://ita.ithea.org>
- April 15, 2010: notification of the paper acceptance
- April 30, 2010: submission of registration forms and visa application forms to [info@foibg.com](mailto:info@foibg.com)

### Papers

Papers from members of ITHEA International Scientific Society (ITHEA ISS) will be preferably included in the events of ITA 2010. Membership of ITHEA ISS is free and may be done by registration at the [www.ithea.org](http://www.ithea.org).

Papers need to be up to 8 pages and formatted according the the sample sheet given at [http://www.foibg.com/conf/paper\\_rules.doc](http://www.foibg.com/conf/paper_rules.doc).

Papers accepted by the Program Committee will be published in several separate numbers of the International Book Series "Information Science and Computing" (IBS ISC).

The best ideas *personally presented* by their authors at the ITA 2010 will be offered a free publishing in English up to 16 A4 pages per idea in the International Journal "Information Theories and Applications"<sup>®</sup> (IJ ITA) as well as the International Journal "Information Technologies and Knowledge"<sup>®</sup> (IJ ITK).

### Submission

Papers need to be submitted via submission system <http://ita.ithea.org>.

Official languages: English and Russian





