

**VARNA FREE UNIVERSITY "CHERNORIZETS HRABAR"  
INSTITUTE OF MATHEMATICS AND INFORMATICS  
OF THE BULGARIAN ACADEMY OF SCIENCE**

**INFORMATICS  
IN THE SCIENTIFIC KNOWLEDGE**

**INTERNATIONAL SCIENTIFIC CONFERENCE**

**2014**

From 27 to 29 June 2014 the Fifth International Conference "Informatics in Scientific Knowledge" - ISK'2014, organized by Varna Free University "Chernorizets Hrabar" and the Institute of Mathematics and Informatics of the Bulgarian Academy of Science was carried out in Varna. The achievements in application of informatics and information technologies in scientific knowledge were discussed in the presented papers. Special attention was paid to the information support in teaching and research work.

**Main Topics:**

- Informatics in economics, management and culture
- Informatics in public administration, law and corporate security
- Informatics in engineering and technologies
- Informatics in education
- Mathematical methods in informatics

The conference was jointly chaired by

**Prof. Dr.Sc. Anna Nedyalkova, President of Varna Free University "Chernorizets Hrabar"**  
and

**Corr. Member, Prof. Dr.Sc. Yulian Revalski, Director of IMI at Bulgarian Academy of Science**

**INTERNATIONAL PROGRAM COMMITTEE**

**Co-chairs:**

Yulian Revalski	Bulgaria
Nikolai Lyutov	Bulgaria

**Members:**

Alexander Polyakov	Ukraine
Alexander Pahomov	Russia
Alexey Nikulin	Russia
Alexey Petrovsky	Russia
Ali Güngör	Turkey
Boris Arkadieievich Skorohod	Ukraine
Dana Petcu	Romania
Jie Lu	Australia
Evgeni Pashkov	Ukraine
Erich Gassner	Austria
Gilles Gonkavles	France
Gyula Katona	Hungary
Ivan Jelinek	Czech Republic
Leon Rothkrantz	Netherlands
Marcello Giacomantonio	Italy
Mile Stankovski	FYR Macedonia
Norvuz Allahverdi	Turkey
Panagiotis Bamidis	Greece
Piet Kommers	Netherlands
Ravil Muhamediev	Kazakhstan
Sergey Andreevich Dubovik	Ukraine
Tanya Kolemishhevska – Gugolovska	FYR Macedonia
Türan Süleymez	Turkey
Franziska Klug	Sweden
Hassane Abuaissa	France
Cvetko Andreeski	FYR Macedonia
Viktors Gopejenko	Latvia
Avram Eshkenazi	Bulgaria
Atanas Palazov	Bulgaria
Galina Momcheva	Bulgaria

Edi Chakarov	Bulgaria
Krasimir Markov	Bulgaria
Krasimira Stoilova	Bulgaria
Lyubomir Sotirov	Bulgaria
Lyudmila Dimitrova	Bulgaria
Neli Maneva	Bulgaria
Radoslav Yoshinov	Bulgaria
Rusalin Nikolaev	Bulgaria
Teodora Bakardzhieva	Bulgaria
Todor Stoilov	Bulgaria

**Organizing Committee**

**Chair:**

Zdravko Slavov	Bulgaria
----------------	----------

**Scientific Secretary:**

Maria Bruseva	Bulgaria
Neli Maneva	Bulgaria

**Organizing Secretary:**

Antonina Ivanova	Bulgaria
------------------	----------

**Technical Secretary:**

Pavleta Krasteva	Bulgaria
------------------	----------

**Members:**

Veselina Spasova	Bulgaria
Evgeniya Rakitina	Bulgaria
Konstantin Kossev	Bulgaria
Krasimir Ivanov	Bulgaria
Nikolay Dimitrov	Bulgaria

## CONTENT

### INFORMATICS IN ECONOMICS, MANAGEMENT AND CULTURE

Crowdfunding Business Models and Their Use in Software Product Development – <i>Todor Branzov, Nelly Maneva</i> .....	7
Software Value Assessment by Means of the Cost approach, as Part of the Company's Intellectual Property – <i>Mariya Bruseva</i> .....	17
Some Cyber Security Issues – <i>Mariya Bruseva, Antonina Ivanova</i> .....	22
Electronic Research on Judicial-Executive Activities – <i>Eddie Chakarov, Kaloyan Iliev, Dimitar Georgiev</i> .....	34
A Prototype of Business Rules Extraction Framework – <i>Delyan Dimitrov, Dimitar Ivanov, Haralambi Haralambiev, Krassimir Manev</i> .....	40
Simulation Model the Dynamics of the System Arising from Interaction of Tourism with the Local Environment – <i>Viktors Gopejenko, Yelena Kozmina</i> .....	53
Protection of Patient's Rights in eHealth Environment – <i>Iskra Mircheva</i> .....	63
Crowdsourcing – a Bridge Between Academia and Commerce – <i>Evgeniya Rakitina-Qureshi, Blagoy Klimov, Gergana Zhelyazkova, Krasimir Ivanov, Ivo Rakitin</i> .....	73
Architecting Synergy Between Technology-Based Innovations for e-Governments and e-Governance to Enable Business Innovations – <i>Alexander Samarin</i> .....	83

### INFORMATICS IN ENGINEERING AND TECHNOLOGIES

Uniqueness of the Image Lighter in Size for Web Space – <i>Maria Deshkova</i> .....	99
Big Data Storage in HDFS and Hadoop in Clouds – <i>Vladimir Dimitrov</i> .....	108

An Analytic Study on the Security Issues of Cloud Computing – <i>Ananthi Sheshasaayee, K.Geetha</i> .....	117
Reduction of Learning Sample Size by Filtering Wells on Correlation Coefficient of Apparent Resistance – <i>P. Gricenko, Y.I. Kuchin,</i> <i>R.I. Muhamedyev, Zh. Nurushev, K. Yakunin</i> .....	127
Using NODE.JS in Developing Web Services – <i>Krasimir Ivanov, Evgeniya Rakitina</i> .....	131
Approaches to Visualization of Apoptosis Models – <i>A. Mansharipova, E. Muhamedyeva, R. Muhamedyev</i> .....	137
Storing Big Data Using Natural Language Addressing – <i>Krassimir Markov, Krassimira Ivanova, Koen Vanhoof,</i> <i>Benoit Depaire, Vitalii Velychko, Juan Castellanos,</i> <i>Levon Aslanyan, Stefan Karastanev</i> .....	147
A Theory and Approach for Developing Cloud Applications – <i>Danko Naydenov</i> .....	165
Geophysical Research Data Preprocessing Module – <i>P. Gritsenko, Y. Kuchin, R. Muhamedyev, Z. Nurushev, K. Yakunin</i> ...	175

## INFORMATICS IN EDUCATION

Programmable Toys and Software for Simulation Work of Robotic Systems as a Tool for Development of Computer Skills – <i>Angel Angelov</i> .....	187
Application of Gamification in University Marketing – <i>Teodora Bakardjieva, Milen Sotirov</i> .....	191
Model for Application of Game-Based Learning in Secondary School – <i>Todorka Glushkova</i> .....	203
Online Document Collaboration Tools in Education – <i>Antonina Ivanova</i> .....	211
Application of Information Technologies to Primary Education – <i>Stefka Kovacheva</i> .....	217



Grading Systems in Teaching of Programming – <i>Krassimir Manev</i> .....	230
MAPLE and MATLAB in Teaching Definite Integral – <i>Zlatka Mateva, Nedka Pulova, Violeta Bojikova</i> .....	240
Against the Boredom Tablet. Get One! – <i>Desislava Milenkova, Yordanka Parvanova</i> .....	253
M-learning in Education – <i>Mitko Mitev, Ivaylo Penev</i> .....	255
Internet and Aggressive Behavior – <i>Joanna Nedyalkova</i> .....	264
Stacks and Queues in Developing Card and Memo Games – <i>Eleonora Pavlova</i> .....	267
Derivative of Function with MAPLE – <i>Nedka Pulova, Zlatka Mateva</i> ...	275
Moodle Gamification – <i>Milen Sotirov</i> .....	290
Adaptability to the Knowledge of the Students by Using of SCORM e-Content – <i>Asya Stoyanova, Todorka Glushkova</i> .....	

### **MATHEMATICAL METHODS IN INFORMATICS**

Translationally Invariant Solutions of the Helfrich Shape Equation – <i>Eddie Chakarov, Vladimir Pulov, Ivailo Mladenov</i> .....	307
Discovering Issues in Code Using Call and Reference Graphs – <i>Todor Cholakov, Dimitar Birov</i> .....	315
Two-Phase Regression with Smooth Transition – <i>Stoyan Nenkov</i> .....	323
Decision Power Indices in Weighted Voting Game – <i>Zdravko Slavov, Christina Evans</i> .....	332
Distance, Similarity and Dissimilarity Measures in Statistical Analysis of Binary Data – <i>Zdravko Slavov, Desislava Slavova</i> .....	346



# **INFORMATICS IN ECONOMICS, MANAGEMENT AND CULTURE**

## **CROWDFUNDING BUSINESS MODELS AND THEIR USE IN SOFTWARE PRODUCT DEVELOPMENT**

Todor Branzov, Nelly Maneva

**Abstract:** *The paper presents some results of our research work on the crowdfunding financial scheme and its possible usage for software projects. In order to investigate the problem we have prepared role models of donations-based and rewards-based crowdfunding and process models of the three main financial schemes– All-or-Nothing, Take-It-All and Stretched-Goals. A linear cost model for fundraisers is presented and its usefulness is commented. Further we have made an assumption that beyond mere financial source, crowdfunding can have significant influence over several key areas of software engineering. We have stressed on several areas and reviewed a possible joint application of crowdfunding and Participatory Design approach. The paper ends with some propositions for future and in-depth research.*

**Keywords:** *Computer Software, Crowdsourcing, Business Models, Donations-Based Crowdfunding, Rewards-Based Crowdfunding, Participatory Design.*

### **INTRODUCTION**

The global software market grew by 6.6% in 2011 to reach a value of \$292.9 billion. In 2016, the global software market is forecast to have a value of \$396.7 billion, an increase of 35.4% since 2011 [5]. Such numbers illustrate the constantly growing importance of software products and services in the global economy and in our everyday life. In the last 30 years the software companies and startups became very attractive for investors and venture capitalists and served as a permanent experimenting ground for emerging new financial instruments and innovative business models.

One such promising financial scheme is Crowdfunding. Generally speaking, it is a form of crowdsourcing focused on obtaining part or all of the necessary financial resources for a project through a massive online public fundraising campaign. Although the campaign is usually “online”, very often some other traditional “offline” techniques and tools are also used.

In Wikipedia an ordered list of top crowdfunded projects can be found, showing that 10 of top 20 highest funded projects of all time (until the end of 2013) are software entertainment products [6]. According to the report on research of Massolution [2], the crowdfunding campaigns performed only through specialised crowdfunding platforms, raised to 2.7 billion USD during the fiscal 2012. About 1.5 billion USD of that money

are gathered without any promises for financial award or return of financial investment to the public investors.

Four models of crowdfunding can be differentiated [4] taking into account the relations between backers and fundraisers:

- Equity-based crowdfunding – backers receive equities in the company;
- Lending crowdfunding – backers receive interest over the sum they pledge;
- Donations-based crowdfunding – backers donate without expecting any material or virtual rewards in return (however, sometimes fundraisers can decide to express their gratitudes);
- Rewards-based crowdfunding – backers receive predefined material or virtual goods based on the amount they donate.

The main goal of the paper is to present the intermediate results of our research efforts in modelling the reward-based and donations-based crowdfunding. We decided to exclude from our study the other two forms - equity crowdfunding and lending crowdfunding due to their venture investment character which makes the decision model for the backers and the cost model for the fundraisers very different.

The proposed role model, process models and the cost model will be further analysed so as to examine their feasibility and usefulness as a financial source for software development and to show how they could influence some activities, recognised as significant during the software life cycle. We try to make parallel with a number of recent achievements in software engineering so as to mark some features of crowdfunding model that reveal its impact beyond mere purpose of financing the project. As an example of mutually beneficial use of a crowdfunding scheme and a contemporary software development approach - developed by us version of the Participatory Design [7], we consider their possible joint application. The investigated hypothesis is that the popularity of the crowdfunding scheme reflects the emancipation of the general public, increases the level of its technical culture and enriches it with some new responsibilities: acting not only as clients, but to participate directly in software development and decision making.

## **A HISTORICAL OVERVIEW**

Since the online payment processor gateways appeared in early 2000's, many enterprises put "Donate" button on their sites. If "the crowd" likes the site or enterprise it can donate for its future development or to support its existence. An example for this is the Wikimedia Foundation, which is doing regular fundraising campaigns for keeping its main project

– Wikipedia, independent and free for all. If the enterprise is with a high profile and there is a perception for its high public value, then the response to such campaign or initiative can raise millions of USD – the Wikimedia Foundation raised 35.2 million USD through 2 million donators in US during the fiscal 2012 and 32 million USD in the fiscal 2013 to that moment [6].

Although this is also a form of crowdfunding, the term is usually used more narrowly for describing an online fundraising campaign for a project, which is in its planning phase and before any result of this project actually exists.

Two kinds of crowdfunding - donations and rewards-based crowdfunding, use a simple idea, successfully explored for ages: a project is announced in public and the author(s) of the idea or organising committee is pleading to the community to support the project. Their request is based on the assumption that the project has significant value for the public in general or for some large social group. A number of institutions like art galleries, churches, universities and schools are funded that way even nowadays.

Crowdfunding uses similar approach and applies Web technologies and Web marketing techniques in its implementation. Especially the donations and rewards-based crowdfunding are very popular in USA, Canada, UK, Australia and South East Asia in the last three years, but are still rare or even unknown in the rest of the world.

Crowdfunding in general is a hot phenomenon with considerable media hype, generated by several projects with tremendous financial success. The most successful use of the crowdfunding financial scheme till now is the MMORPG software project, named “Star Citizen”, which announced about 40 million USD fundraised with rewards-based scheme in March 2014. Also in March 2014 the leading crowdfunding platform “Kickstarter” operates for 762 projects tagged as software and 4362 projects tagged as video games.

#### **A ROLE MODEL FOR DONATIONS AND REWARDS-BASED CROWDFUNDING**

Three main roles can be differentiated in donations and rewards-based crowdfunding fundraising campaign process – Fundraiser, Backer, Mediator. The role of the Mediator is optional and includes all third parties that can have direct influence over the process – most notably crowdfunding Web sites and online payment processors. The Mediator role in the model is designed to be complemented with the other third parties if and when they appear in the future.

We propose the following role model, describing the main roles and the relations between them in donations and rewards-based crowdfunding (see Fig.1).

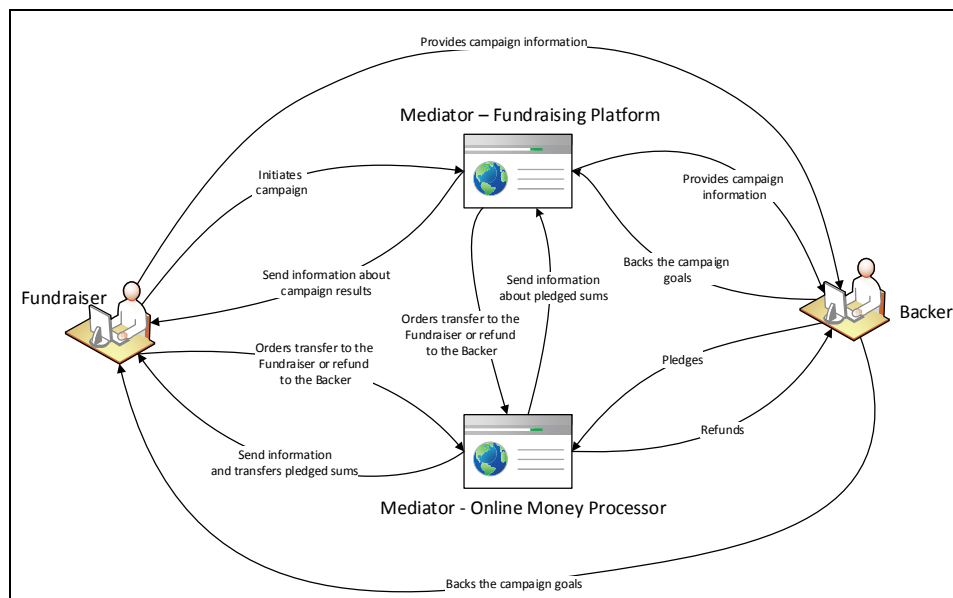


Figure 1. Crowdfunding role model

The main responsibilities, which we have identified for the Fundraiser's role are:

- Initiates a project;
- Announces the project in a fundraising platform or/and in its own site;
- States and sets fundraising scheme;
- Completes some activities needed to provide all necessary campaign information to the potential backers;
- Transfers back pledged sums to backers (depends on chosen crowdfunding process model).

The main responsibilities, which we have identified for the Backer's role are:

- Gets information about the project;
- Pledges a sum;
- Gets satisfaction or some reward;
- Can be refunded if the stated goal is not achieved (depending on the chosen crowdfunding process model)

The main responsibilities, which we determine for the Mediator's role, accomplished through online fundraising platforms and online payment processors are:

- Serves as intermediate in all money transactions and information for pledged sums;
- Guarantees refund to the backers if stated goal is not reached (depends on the chosen crowdfunding process model)

## DONATION AND REWARDS-BASED CROWDFUNDING PROCESS MODELS

Our study shows that three donations and rewards-based crowdfunding process schemes are dominant at the current moment, namely All-or-Nothing (AoN), Keep-It-All (KIA) scheme and Stretched Goals Scheme (SGS).

Both AoN and KIA schemes include setting a fundraising goal (minimal amount that have to be raised), but in AoN scheme if that goal is not reached, all backers receive their pledges back. In KIA scheme if the goal is not reached, the backers can receive back their pledges by discretion of the Fundraiser.

In SGS scheme several stretched goals are stated and bonded with the declaration of the Fundraiser to add some pre-specified additional value to the product or the service. If and when the stretched goal is achieved, the Fundraiser takes all needed actions to add that value.

We have used the swim lane diagram with basic flowchart notation to propose a model of these three schemes within the previously described role models of donations and rewards-based crowdfunding (see Fig.2).

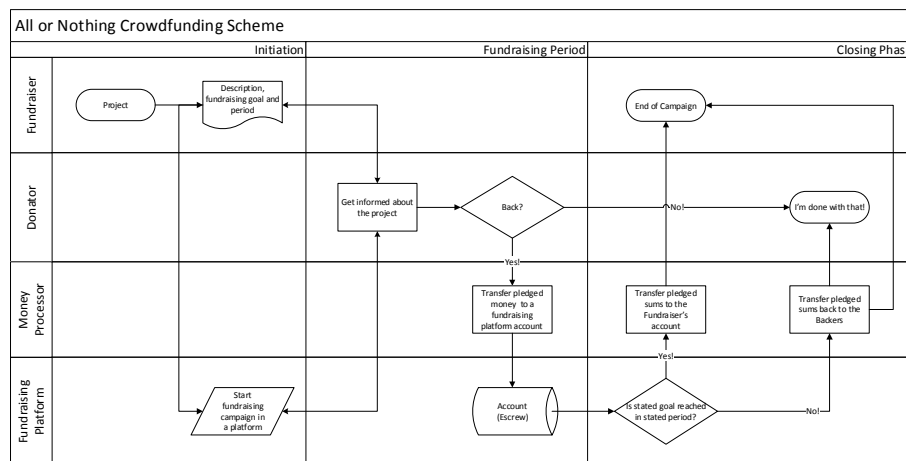


Figure 2. AoN swim lane diagram

In the AoN process description we assume that the fundraising platform is used by default. In its role as a Mediator, the fundraising platform guarantees that the backers receive their pledges back, if the stated fundraising goal is not reached. We assume that AoN scheme can be more effectively applied when the Mediators are active – by definition even if a fundraising platform is not used, the Backers have to receive their pledges back if the stated fundraising goal is reached, but the fundraising platform guarantees that to happen in AoN scheme, and we assume this as one of the main reasons for the popularity of the fundraising platforms. Also if a Mediator is used he/she takes part of the administrative burden from the Fundraisers and have to use more resources in the AoN scheme. If a fundraising platform is not used at all, then the AoN process model is a subset of the KIA process model.

In KIA process description (see Fig.3) we state that fundraising platform is an optional (shown by the dashed connection lines between the related activities). If a fundraising platform is not used, the KIA process model and AoN process model are identical.

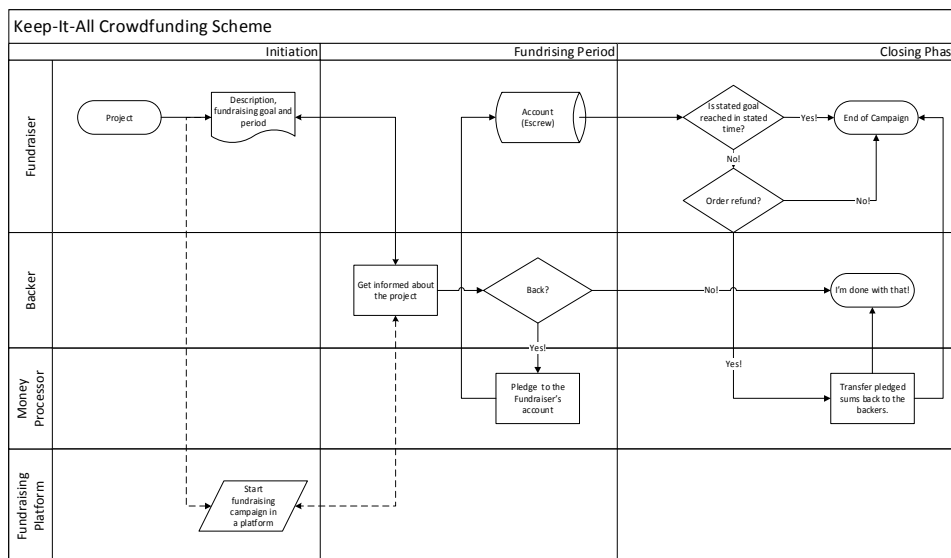


Figure 3. KIA swim lane diagram

The SGS scheme is recognised by some of fundraising platforms (i.e. Kickstarter) but there is no platform that has functionality that allows setting a SGS campaign. Thus the proposed SGS process model describes fundraising platform only as an information and advertising tool (see Fig.4).



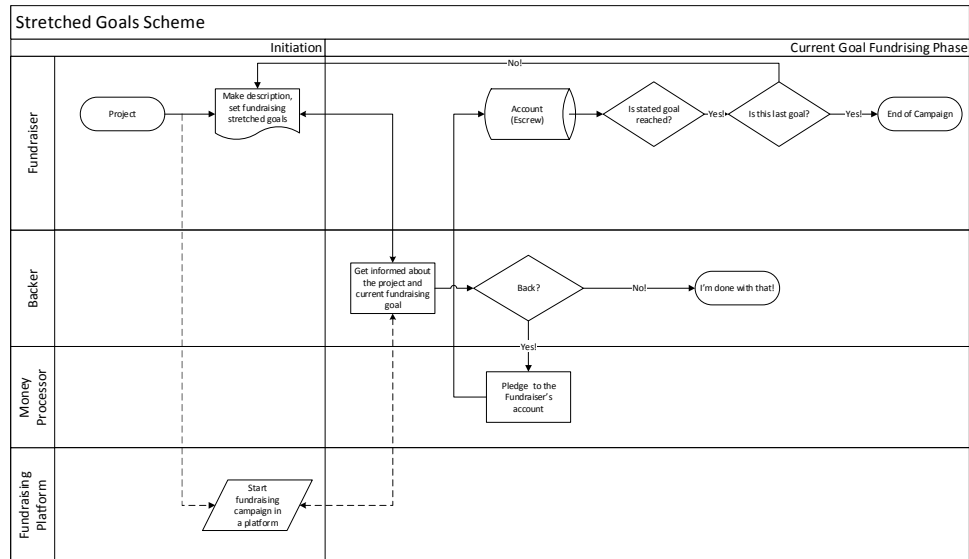


Figure 4. SGS Swim Lane diagram

Studying the crowdfunding process models and observing the number of key activities in Mediator's swim lanes we can definitely conclude that the Mediators have more influence in the AoN scheme than in any other.

### A LINEAR COST MODEL FOR FUNDRAISERS

To expand the suggested role and process models we define and propose the following linear cost model, evaluating the actual cost of the investment to fundraisers.

*Rewards based = cost of announcing and advertising + cost of fundraising platform + cost of online money processor + cost of administering campaign + sum(cost of rewards in packages).*

*Donations based = cost of announcing and advertising + cost of fundraising platform + cost of online money processor + cost of administering campaign.*

Applying the above model, one can also use a rule of a thumb that the cost of the fundraising platform is about 4% of the sum and also the cost of the online money processor is about 3% of the sum.

## **CROWDFUNDING IN SOFTWARE PRODUCT DEVELOPMENT PROCESS**

Crowdfunding is one of the emerging business models in the software industry. We assume that crowdfunding can go beyond mere business process and that “the crowd” actually can have significant influence over the whole software engineering process, especially over requirements gathering and quality assurance.

In every rewards-based crowd funded software project we have studied, the rewards always included some form of early access (before the official market release) to the product. As every beta tester is obliged to sign some form of agreement that he/she will report bugs and the developers give them tools to do that – this early access actually provide for developers a team of functionality testers. This team is so highly motivated that every member is eager even to pay for the opportunity to be a tester.

Also some of the projects in SGS scheme gave an opportunity to the backers to pledge to several packages of functionality or content simultaneously, effectively giving them a way to influence the prioritization of the features development.

We suggest that in every agile software development lifecycle crowdfunding can easily be a source of the so needed “involved client”. Also, the SGS scheme in particular, is a perfect match for an agile lifecycle as the both are based on the idea of many relatively short periods of time at the end of each some clear and distinct additional value is added to the constantly evolving product.

We also suggest that crowdfunding can be used along with a software development approach of Participatory Design (PD). The original model of PD describes user involvement in the design stage of a classical lifecycle only. A modification of this approach has been described in [7], showing how it can be used for usability assurance. Also it was proposed that the development technique of “evolutionary” prototyping can be applied to PD. It suggests that a sequence of “partial” software products is produced, each of them being an extension of the previous one. When a prototype is ready the user has to test it in a real-life environment and to give a proper feedback to the developers. With crowdfunding rewards packages that include early access to the product, the users effectively are involved in the process. In aforementioned “Star Citizen” project the rewards packages included some forms of early alpha-stage access – initially the users had access only to a 3D-model of a hangar, the initial version of the user interface (UI) and only one 3D-model of a spaceship, no other functionality. As the “prototype” evolved, and users gave their suggestions, the UI got redesigned accordingly.

## **CONCLUSIONS AND FUTURE WORK**

Realizing the constant growth and social impact of the software market, we are studying a modern technique for financing some creative and software projects – crowdfunding. It facilitates the start of a project, makes a public recognition of its social value and gives a sense for participation in the development process to the public. The purpose of this paper is to give a broad view on the topic and to mark some areas for future in-depth research.

We suggested that two crowdfunding schemes - donations and rewards-based, have to be perceived differently than others in existence today as they do not offer any financial returns-of-investment to the investors. We proposed several models: a role model, describing the different roles and relations in the donations and rewards-based crowdfunding schemes; three process models and a cost model for the fundraisers. These models can be used in the future research on that topic.

We made an initial research on the benefits of unifying the crowdfunding scheme with some contemporary software development techniques, e.g. agile techniques and the special participatory design.

Our future work on that topic can be in the following directions:

- To gather more empirical data and use them for improvements of the proposed models;
- To apply some formal methods to validate and assess the models;
- To investigate further the possible integration of the crowdfunding with some other software engineering activities;
- To study the psychological profile of the Backers in order to find out how to increase their motivation and desire to influence the software development process and its quality results.

## **ACKNOWLEDGEMENTS**

This work is supported by the National Scientific Research Fund under the Contract ДТК 02-69/2009.

## **References**

- [1] Klæbe, Helen G., Laycock, Rebecca. How to work the crowd: a snapshot of barriers and motivations to crowdfunding. Australia Council for the Arts, 2012.
- [2] Maclellan, Kylie. Global Crowdfunding Volumes Rise 81% In 2012. Reuters, 2013.
- [3] Hernandez, Megan. Fundraising 2013. Wikimedia Foundation, 2014.
- [4] Rothler, David, Wenzlaff, Karsten. Crowdfunding Schemes in Europe. EENC Report, 2011.

- [5] Software: Global Industry Guide. MarketLine Report, December 2012  
[6] List of highest funded crowdfunding projects. Wikipedia, 2014.  
[7] Maneva, Nelly. Participatory Software Development as a Basis for Usability Programs. Proc. of the CompSysTech '2002, II.6-1, II.6-6.

**Authors**

**Todor Branzov**, Institute of Mathematics and Informatics of BAS,  
+359(88) 5095757; [tbranzov@math.bas.bg](mailto:tbranzov@math.bas.bg)

**Nelly Maneva**, PhD, Professor, Institute of Mathematics and Informatics  
of BAS, [neman@gbg.bg](mailto:neman@gbg.bg)

# SOFTWARE VALUE ASSESSMENT BY MEANS OF THE COST APPROACH, AS PART OF THE COMPANY'S INTELLECTUAL PROPERTY

Mariya Bruseva

**Abstract:** *After analyzing the intangible assets of the particular company in the proposed work are considered used modern approaches in respect of determination of the fair price. Attached is a specific method for database evaluation-element of the intellectual property of a company.*

**Keywords:** *intangible assets, approaches to valuating firm, market price, cost approach, assessment of copyright*

## Introduction

By definition, the business cost as much as potential investors are ready to pay for it. Depending on the specific business objectives, evaluation purposes, factors influencing the assessment and the market situation, the valuation of a business can vary significantly. The company market position is also an important assessing factor. Assessment can be greatly affected by the company market share, its trade mark recognisability, loyalty to its business and brand, copyright, patents and more. Company assessment is significantly affected by its ability to generate positive future cash flows. Most commonly used internationally recognized methods for business evaluation are based on evaluation of the individual elements in the composition of the assets of a company and their quantification.

After analyzing the intangible assets of the particular company in the proposed work are considered used modern approaches in respect of determination of the fair price. Attached is a specific method for database evaluation-element of the intellectual property of a company.

## 1. Role and place of intangible assets in the company's development

With the market relations development a new type of resources in the property composition appeared - intangible assets as part of the fixed assets of the company [2], [3].

Following assets can be recognized as intangible assets (IA):

- Ø recognizable and have no tangible (physical) form;
- Ø used in the undertaking;
- Ø that may result in future economic benefits;
- Ø whose useful life is more than 12 months;
- Ø which value can be measured with sufficient reliability;
- Ø documents availability confirming the copyright holder.

In the absence of any of the above criteria expenses are not recognizable as made for intangible assets and respectively are recorded as company expenses.

## **2. Types of Intangible assets and their role in the enterprise development**

Under IAS (international accounting standard) and NSS (National Accounting Standards), intangible assets are internally working intangible resources involved in production and belonging to the enterprise [4]. Intangible assets have no physical characteristics but are of value in terms of advantage or exclusive rights and privileges granted to them in the production process. These include:

- Ø intellectual property;
- Ø right to lease;
- Ø copyright;
- Ø patents;
- Ø the price of business.

Intangible assets are when company spend funds or incur liabilities, on the acquisition, development, maintenance and enhancement of intangible resources such as scientific or technical knowledge, design and implementation of new processes or systems, licenses, intellectual property rights, market knowledge and trademarks (including brand names and publishing titles).

Common examples of items covered by this wide range of definitions are: computer software, databases, patents, copyrights, motion picture films, customer lists, rights to service the mortgage, fishing licenses, import quotas, franchises, relationships with customers and suppliers customer loyalty, market share and marketing rights.

## **3. Approaches to value intangible assets**

In adjust global valuation practice the following approaches to value intangible assets are used [1]: cost, market, income

### **3.1. Cost assessment approach**

In the cost approach calculates the cost of reproduction. When using it, intangible assets are measured as the sum of the cost of acquisition and commissioning. Elementwise assessment is the main feature of the approach, meaning is that intangible assets are divided into parts, followed by assessment of each part and obtaining the entire value of an intangible asset as sum of the value of its parts.

### **3.2. Market valuation approach**

Market approach is used to estimate intangible assets market value from the data of the recently concluded deals with similar intangible assets. The same can be used for these types of intangible assets, for which often transactions are made in the market. Similar items sales prices are the input data for calculation. Substitution is basic principal in this approach, which states that a reasonable investor would pay for the site greater than the value of sales similar object contributing the same

benefits as an object. Therefore, the selling prices of similar objects serve as input data for the cost calculation of the intangible asset evaluations.

### **3.3. Valuation income approach**

The income approach or the approach based on income is widely applicable in the well-developed market economy. Income based approach main advantage is that its methods are directed to the very sources of the value of the assets being evaluated - the future income from their use. Future effects in particular with this approach are the basis for determining the theoretical price of an intangible asset. In this respect we can say that the methods of the approach based on income in largely report the objective of any investment - obtaining a higher value in the future.

### **4. Assessment of copyright (SOFTWARE) by the cost approach**

1. Date and place of assessment: 01.01.hhhhhg., Varna (course - xxx EUR).

2. Objective assessment: inclusion of intellectual property (IP) in the company authorized capital.

3. Object of assessment: IP is a database designed to facilitate and accelerate the necessary information search by the person working on the Bulgarian market for intellectual property.

4. Market condition

The following are the main elements of the development costs (ISD): programmer salaries (with allocation of the budget), the rent payment, equipment amortization payment, overhead and the entrepreneur profit.

An experienced programmer monthly salary, depending on qualifications ranges 500-1000 per month. Office space rent is at a level of 300 to 1000 EUR per year per m<sup>2</sup>.

Professional programmers working with a set of equipment (computers, printers, scanners, etc.), that cost 1200 - 2400 EUR for a period of effective use 3-5 years.

5. Evaluation procedures (cost approach)

To calculate the value of the object is appropriate to apply the method of cost replacement. Therefore, it is necessary to estimate the cost assuming that IP is currently in process of establishing in the framework of a well-functioning organization.

There are two experts' opinions:

- The database can be created by an experienced programmer with an average rating for three months;
- This software can be developed by a highly skilled programmer within one month.

Accordingly, we construct two options for cost calculations

Costs elements (B EUR)	I variant	II variant
<b>Salary: (without taxes)</b>	500	
- one month (with taxes)	700	1500
- three months	2100	-
<b>Total:</b>	<b>2100</b>	<b>1500</b>
<b>Rent for premise</b>		
- Area: 10 m <sup>2</sup>		
- Area (EUR per m <sup>2</sup> per year)	300	600
- Rent for month	250	500
- Rent for threemonth	750	-
<b>Total:</b>	<b>750</b>	<b>500</b>
<b>Amortization of computing</b>		
- Price of set of equipment (computers, printers, scanners, etc.)	1500	2520
- Duration of effective use	3 years	3 years
- Depreciations for one month	-	70
- Depreciations for three months	125	-
<b>Total:</b>	<b>125</b>	<b>70</b>
<b>Total three elements</b>	<b>2975</b>	<b>2070</b>
Overheads - 20%	595	414
Total with overheads	<b>3570</b>	<b>2484</b>
Entrepreneur's profit - 25%	892,5	621
<b>All:</b>	<b>4.462,5 EUR</b>	<b>3.105 EUR</b>

By this way, we get two possible values calculated by replacement cost database 4 462,50 EUR and 3 105,00 EUR. Since this product was created at the end of the year XXXX, we can assume that the reduction in value due to aging has not yet occurred.

#### 6. Results concordance

Since there is no reason to give more weight to some of both options, the final evaluation of the project costs (PV), should be calculated as an average of variant I and variant II.

$$PV = (4\ 462,5 + 3\ 105) : 2 = 3783,75 \text{ EUR}$$



## 7. Conclusion

The estimated database market value is objectively range from 3 000EUR to 4500 EUR.

By the date of valuation, according to experts, it is equal to 3783,75 EUR

## Conclusion

The intangible assets valuation is relatively independent area of current scientific research, which is in constant evolution and needs further perfection. The intangible assets are subject to corporate property, which you can not only own, use and manage, but also (if properly documented and evaluated) you can use as authorized capital in the economic activity of the enterprise. For example, intellectual property in the form of inventions protected by existing patents - "patent pending", "know-how", "brands" etc. widely in the world is highly valued and is a source of income and profitable investment. It is well known that 85% from total cost of the public companies in the developed countries is intellectual property [5]. The market price impact of the rest of the company fixed assets is not to be underestimated. By the proper disclosure and assessment of the intangible assets of company would increase the market value, which will be step towards enhancing its competitiveness and would have a positive sustainable development in the respective sector of the economy.

## References

1. Славов, Здр., Брусева, М, Къновски, Р. Анализ на пазари и вземане на икономически решения в условия на неопределеност. ВСУ, Варна, 2012
2. Иван Златков. Коментар на международните счетоводни стандарти. Изд. Ciela, 2009
3. Иван Златков, Счетоводство, изд. Ciela, 2009
4. Списание "CIO", бр.4/2009
5. <http://www.online-trading-stock.org/valuation-of-intangibleassets/bg>

## Author:

Assoc. Prof. eng. math. **Mariya Nikolova Bruseva**, PhD  
Varna Free University, Chayka resort, Varna, 9007  
Tel: +359 52 359 572, Mob. +359 897 80 55 02  
[mariya.bruseva@gmail.com](mailto:mariya.bruseva@gmail.com)

## SOME CYBER SECURITY ISSUES

Mariya Bruseva, Antonina Ivanova

**Abstract:** *Implementing information security in all business levels, there is a need to understand and summarize different types of attack on computer systems and networks. The paper observes actual security threats like WiFi-viruses, heartbleed, Trojan horse attacks. The 2013 Data Breach Report illustrates some aspects of reasons and results of data breaches and other network attacks. The most significant cyber issues in organizations are analyzed and cyber security challenges in 2014 are displayed.*

**Keywords:** *IT security, cyber security, security threats, cyber attack, cyber issues, data breaches*

### Introduction

By entering the second decade of the 21st century, the world becomes a place where messages are exchanged instantaneously, news transmitted by global media networks with a single click

In a world of global network information is the key to competitive advantage. Information however may become a danger weapon when by accident or deliberate methods ends up in the hands of the wrong person. In such circumstances same information can lead to a threat of critical level for the organization. Information security means protecting information and its supporting infrastructure from accidental or intentional impacts of natural or artificial nature, which can cause unacceptable damage to the owners or to the users as well to the supporting infrastructure. Protect information –means to build complex of measures aimed to ensure information security [2].

Information security is one of the most important aspects of the comprehensive security at any level - national, branch, corporate or personal. That why protection against unauthorized access to information resources , ensuring the security of information and telecommunication systems are one of the most important component of the national interest in the information area of each country. This is the reason why for all economy sectors is necessary possible attacks to be known, what are the preventive methods and how this can be realized.

### 1. Security Systems in modern company

*“From all possible troubles happens this that leads to the greatest damage”*

*Third consequence of  
Murphy's Law*

Unlike all other machines created by humans, computers do not convert energy from one form to another, but data into information. Particularly information is the most valuable component in them.

According to the United States FBI whenever is "pierced" computer security of a large company , it loses an average of 700,000 dollars , and banks losing 8000 dollars per second of crashes and failures of computers. Credit card fraud cause lost about \$ 5 billion annually. Evidence shows that even the user has nothing to hide there always will be someone who willing to turn your computer into "trash heap" as long as this opportunity exist.

To the question "Which system resources should be protected and at what price ?", answers are not always obvious. Intruder who compromises or undermined particular computer usually gets access to its all resources. Some may be interested in the outgoing links of the attacked system only as a "springboard" to attack other, more interesting targets. Protection methods that are sufficiently effective against attacking by modem teenagers are powerless against staunch professional crackers. Another issue is the security cost - there is a fragile balance between too much and too little security.

Security is the ability of a system to withstand internal or unstable factors which can lead to the undesirable condition or its behavior. This also applies to the information systems security.

The purpose of any information system is to provide complete, accurate and timely information. This information is vulnerable because of casual and malicious destabilizing factors (threats) so measures have to be taken for its protection.

## **2. Types and means of attacks on computer systems and networks**

The attack is deliberate action by deliberate infringer, attack may consist a searching of weaknesses in order to collapse the security of information system. An successful attack can achieved:

- ✓ information leaks ;
- ✓ Denial of Services ( lock ) ;
- ✓ external misuse of company resources , theft of services;
- ✓ recording and use the traffic network of the company by outsiders ;
- ✓ information destruction ;
- ✓ data fraud ;
- ✓ harmful programs installation;
- ✓ indirect ( indirect ) abuse (use of other systems to create malicious programs);
- ✓ password breaking;
- ✓ poor administration.

### 3. Actual security threats

*“Each solution delivers new problems”  
Seven consequence of  
Murphy's Law*

The choice of specific measures and tools, and the development of appropriate mechanisms for physical and technical protection are part of the project for the overall protection system. Unfortunately, current technology does not allow for error-free programs, respectively, does not promote the rapid development of tools to ensure information security. Increased number of attacks is not the biggest trouble. Discovering new software vulnerabilities is the worse and as a consequence there are new types of attacks. For example, according to the reports by the National Center for Infrastructure Protection in the United States are reported problems with the software, leading to medium or high risks in its use. Among the "victims" operating systems are almost all as the new systems errors are immediately used by malicious.

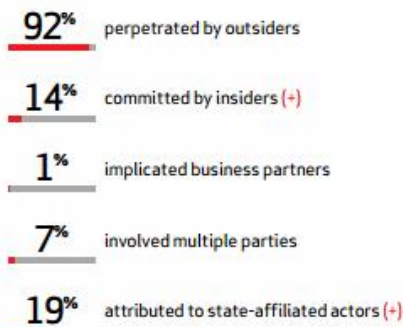
Construction of the data protection concept should be carried out in accordance with prior studies on the vulnerability of the system to specific external influences. For this purpose it is necessary to classify the factors which are able to disrupt the functioning of the systems.

For the correct solution it is necessary the company security treats to be considered.

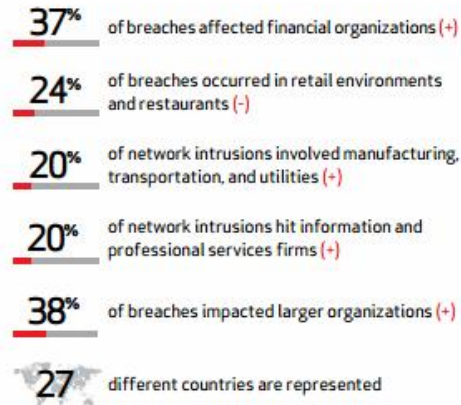
#### **3.1 Data from the report “The 2013 Data Breach Investigations Report”**

The report is based on an analysis of 47 thousand publicized IT security relating incidents, where there are 621 confirmed information leaks and 44 million compromised records [1]. The covered information is for 27 countries from all continents, including Bulgaria. Here are some of the most interesting data:

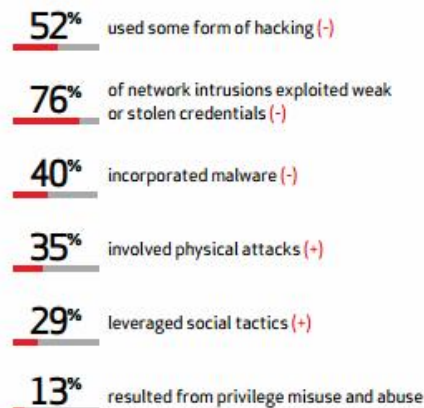
### Who's perpetrating breaches?



### Who are the victims?



### How do breaches occur?



### What commonalities exist?

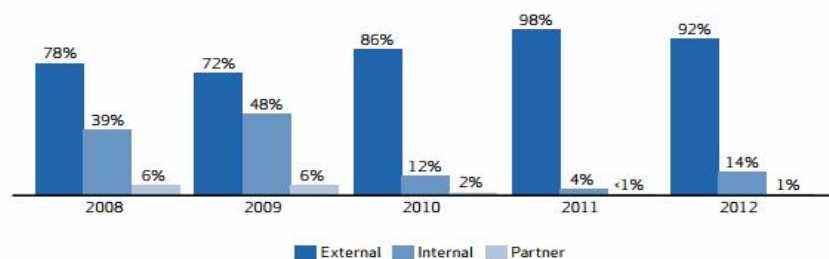
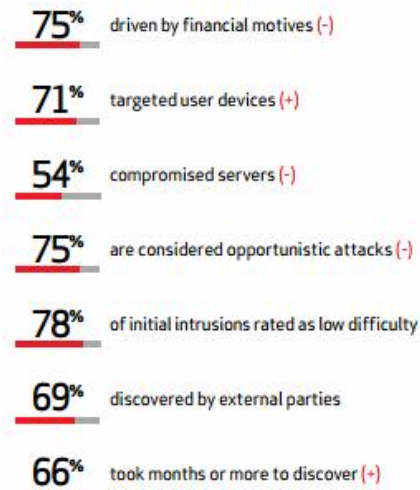


Figure 1. Threat actor categories over time

### **3.2 552 million people suffer from hacker attacks in 2013**

According to the Symantec Company, in 2013 552 million Internet users suffered from hackers attack. The network of the Company Global Intelligence Network, which unites more than 69 million sensors in 157 countries registered significant malicious activity increase on the network, and increase the data leaks incidents in 2013 by 62%.

### **3.3 Study: 'Cleaning' after cyber attack cost average 8.9 million dollars.**

According to Ponemon Institute survey on 56 corporate and government organizations, the biggest expenses to cover the costs associated with cyber attacks are in the United States spend an average amount of 8.9 million dollars in 2012 [5]. This is 6% higher than the results from 2010.

For 2013 the survey goes beyond the United States, comprises the UK, Germany, Australia and Japan. Allocated to cyber attacks cost outside U.S. amounted to a smaller amount - 5.9 million dollars in Germany and 5.1 million in Japan, for example.

"We found that it is most likely that U.S. companies become victims of the most expensive types of cyber attacks like: malicious agents malware and web-based attacks," the report said.

In the UK and Australia, where cybercrime is worth 3.2 and \$ 3.3 million respectively, the most common attacks are the so-called denial-of-service ("refused service"). Least likely to be attacked by malicious code and denial-of-service had Germany, and the low probability of malicious agents and web-based attacks - Japan.

The study identified five external costs factors of cybercrime: business interruption, loss or theft of information, loss of revenue, faulty equipment, and "other." Internal cost factors are: detection, investigation and escalation, retention, recovery and subsequent efforts to repel future attacks.

### **3.4 Fishing 2013: a smaller volume, more precise targets**

According to the (APWG)division company Websense, the volume of phishing attacks in the past year has decreased by 1.12% compared to 2012 and now Phishing messages are not more than 0,5% of electronic letters [6].

Moreover, in 2013 a significant decrease in the number of large-scale phishing attacks is noticed. This is probably due to the fact that hackers are re-direct their resources to other activities - such as DDoS attacks.

Meanwhile Phishing is becoming more purposeful and cleverly, and that means that the majority of users opening attachments or selected

links and by that way falling into their traps. In many cases Phishing messages are sent along with legitimate documents which way hampers their detection by use of reputation technology. Frequently are used personal messages, false reports of inability to deliver mail, invitations to join social networks and more. Still popular approach for hackers is to register fake domain called near popular brands and then use these domains for phishing.

Leader in sending phishing messages remains China. Following China is U.S. and further Germany , UK, Canada , Russia, France , Hong Kong , Holland and Brazil.

### **3.5 'Trojan horse' attack financial institutions**

*„Rob banks because there is money“*

*Willie Sutton  
bank robber*

They say that Mr. Sutton has not responded by that way, but this quote is still relevant. Paradigm is updated in particular when it comes to today's malware for financial institutions. Money circulated in the online banking process explaining why same are in the focus of the attacks. We should not be surprised by the fact that we observe a steady development in Trojans targeting online banking services. One example that recently wrote a Symantec 's blog about Trojan Neverquest, successor of the Trojan. Snifula, which first appeared in 2006 and is still in use [7].

During the first nine months of 2013, the number of infected, as result of the work of the most common types of Trojan horses grew by 337% . These are almost half a million susceptible to fraud infected computers within a month. For a better understanding of the mechanics behind the Trojans in financial institutions and the limits of their capabilities, Symantec experts have analyzed more than 1000 modern configuration files belonging to 8 Trojans hijacking data services for online banking. These configuration files define which URL should attack Trojan and which attack strategy to use. Cyber attacks range from a mere reference to a forged website to sophisticated techniques called Web-injects, which can automatically perform transactions and siphoned money. The analyzed configuration files are targeting a total of 1486 organizations. This proves once again the widespread use of Trojan horses targeting anything that can bring monetary benefit for cyber criminals.

Most often attacked bank are located in the United States, it is present in 71.5% of all analyzed configuration files with Trojans. All the 15 most often targeted banks are present in more than 50 % of the

configuration files. This means that every second Trojan targets at least one of these banks. These high values may be due to the fact that the targeted URL addresses appear as examples in some basic manuals sold with Trojans. Another possible reason could be that the Trojans continue to act against these companies since not all financial institutions have switched to a higher level of authentication and security.

Of course, most financial institutions are aware of the development of cyber crime and develop new defense mechanisms to block this type of cyber attacks. Unfortunately, the new security measures implementation require time and money, and cybercriminals will always find new ways to attack. Symantec expectations are that in 2014 will continue to monitor cyber attacks targeting online banking services.

### **3.6 Wi-Fi viruses - a new threat to IT security**

Researchers from the Liverpool University demonstrated that wireless Wi-Fi networks can be infected with a virus that in densely populated areas is distributed with no less efficiency than influenza viruses. Information on this topic was published in EURASIP Journal on Information Security.

### **3.7 Heartbleed**

Heartbleed is the most serious threat to IT security and is probably one of the most serious bugs in the Internet security over the past five years. Error in OpenSSL allows to any attacker to infiltrate your computer and steal whatever he wants without leaving a trace [8]. This can happen if you accidentally visited a website hosted on a server that used OpenSSL old code. The problem is that until a few weeks ago on the net there was hundreds of thousands of similar pages.

IT Security experts advise users to change their passwords after receiving assurance from their ISP that the necessary changes to the server has been made. They assure us that after Heartbleed was identified, the Internet is considerably more secure.

## **4. Cyber security challenges in 2014**

New trends in IT pose new challenges to managers and professionals involved in the information protection. As it was noted earlier safety is an important aspect of IT, which affects every component of IT solutions - from sensors and standard equipment to key systems in large corporate data centers. Whatever information technology to use , you should always think about safety and have an action plan in case it proves to be compromised - especially in today's world of hyper connectivity.



Enterprise IT systems have come a long way - from the first steps on the Internet when connect to online services, most users use short and easily memorable passwords, until today , when two and three-factor authentication and so called "Encryption from end to end" are used. Currently , among the important trends in the development of IT in terms of safety is the increasing number of portable computers, more active use of wireless connectivity with short-range (i.e. Near field communication, NFC), and also the growing number of home devices connected to networks. Many analysts agree that the above trends pose new tasks related to IT security. The publication authors discuss the basic expectations of the Crawley Company that are similar to the expectation of many other experts in this field. [4]

**Experts predict that heightened expectations among consumers, advocates and regulators will require organizations to step up cyber measures in the year ahead**

Kroll, the global leader in risk mitigation and response solutions, today released its third annual Cyber Security Forecast, a prediction of the most significant cyber issues organizations will confront in 2014. The latest forecast highlights seven trends identified by Kroll and suggests that a changing tide in cyber standards, both social and legal, will require organizations to take stronger actions and safeguards to protect against reputational, financial and legal risks [3].

**1. NIST and similar security frameworks will become the de facto standards of best practices for all companies.**

Between the Snowden fallout and increased scrutiny by the FTC and other regulators in the U.S. and around the world, cyber security strategies that were largely designed for companies that were part of the "critical infrastructure" will become more of an expectation for everyone, from conducting an effective risk assessment to implementing sound cyber security practices and platforms. Whether compulsory or unstated, these standards will drive organizational decision-making with regard to cyber security. Organizations that don't follow suit may find themselves subject to shareholder lawsuits, actions by regulators, and other legal implications.

"This trend will move the U.S. in the direction of the EU, where there is a greater recognition of privacy as a right," said Alan Brill, senior managing director at Kroll. "As new laws evolve that reflect the NIST guidelines and look more like the EU privacy directive, some U.S. companies will find themselves ill-prepared to effectively respond to the regulations. To minimize their risk, organizations will have to get smart on these standards and make strategic business decisions that give clients and customers confidence that their information is protected."

**2. The data supply chain will pose continuing challenges to even the most sophisticated enterprises.**

It is not unusual for companies to store or process the data they collect by using third parties. However the security that these third parties use to safeguard their client's data is frequently not understood until there is a breach. Additionally, companies may believe that their subcontractors will notify and assist them in the event of a breach. Unfortunately, this is often not the case. Companies will need to vet their subcontractors closely and get specific as to the technical and legal roles and responsibilities of their subcontractors in the event of a breach.

"Kroll has responded to breaches where subcontractors not only failed to provide timely notice that they were breached, but also refused to cooperate with the investigation. Companies should know who they are giving their data to and how it is being protected," said Tim Ryan, managing director and Cyber Investigations practice leader. "This requires technical, procedural, and legal reviews."

**3. The malicious insider remains a serious threat, but will become more visible.**

"Et tu, Brute?" Whether it was Shakespeare's Caesar or America's Benedict Arnold, we have long known the pain of betrayal by those we trust. Information technology simply made the betrayer's job easier. In 2014, a significant number – if not almost half – of data breaches will come at the hands of people on the inside. However, as the federal government and individual states add muscle to privacy breach notification laws and enforcement regimes, the hidden nature of insider attacks will become more widely known.

"There's a tremendous amount of data compromised today where the act is never discovered or disclosed. People discount the insider threat because it doesn't make the news. Instead, we see headlines about external credit card breaches and theft of personally identifiable information, because regulations mandate accountability and punishment is expensive. The insider threat is insidious and complex. Thwarting it requires collaboration by general counsel, information security, and human resources. SEC breach disclosure of "material losses" may be the model for rules requiring a company to be more transparent and answerable for allowing bad actors to go unpunished," said Ryan.

**4. Corporate board audit committees will take a greater interest in cyber security risks and the organization's plans for addressing them.**

With more and more data breaches – from theft of trade secrets to loss of customer information – in the headlines, corporate audit committees are beginning to focus on the connection between cyber security and an organization's financial well-being. As such, they will

expand their attention beyond the financial audit process to the organization's strategic plans for protecting non-public information and risk mitigation plans for responding to a possible breach. CIOs and IT leadership should prepare accordingly.

"Organizations recognize that it's their duty to protect against the loss of information and its associated risks," said Brill. "As corporate boards carry out their fiduciary responsibilities, they must also protect the company from possible shareholder lawsuits that allege the company's cyber security wasn't at a level that could be reasonably viewed to be 'commercially reasonable' and that incident response plans weren't in place to mitigate the risk. The challenge they face is determining what is a reasonable level of security and response, and who should make that call – is it their IT team, an industry expert, an independent third party?"

**5. Sophisticated tools will enable smart companies to quickly uncover data breach details and react faster.**

Companies realize that even the best firewalls and intrusion detection systems cannot stop all attacks; the most secure firms experience computer security incidents. But technological progress over the last 12 months will enable companies to unravel events and see with near-real-time clarity what's happened to their data and how much damage has been done. That is, if companies choose to change.

"Most organizations have invested in preventative security technologies, but remain unprepared to launch an effective response to a leak or intrusion. Without the right tools and policies in place beforehand, they find themselves suddenly under intense pressure to investigate, track, and analyze events," said Ryan. "It takes more money and time to scramble at the last minute. We've seen a dramatic improvement in response technology over the last year. Companies have never had a better opportunity to enhance their existing protocols with a methodology that can mean an informed and timely response. There's no reason not to be prepared."

**6. New standards related to breach remediation are gaining traction and will have a greater impact on corporate data breach response.**

Credit monitoring will no longer be the gold standard in breach remediation in 2014, as lawmakers, consumer advocates and the public at large continue to raise questions about the relevancy and thoroughness of this as a stand-alone solution and demand a more effective alternative. While no legal guidelines currently exist for consumer remediation, the FTC and states like California and Illinois are already offering guidance that suggest a risk-based approach to consumer remediation – one that matches remedy to individual risk

based on the unique circumstances of a breach – will be the way of the future.

“The notion that credit monitoring is a panacea for all data breaches is misguided. When you couple the myriad types of sensitive information with the multitude of ways an identity can be stolen and used fraudulently, there are many instances where credit monitoring will not be helpful to a breach victim at all, including medical identity theft, criminal impersonation, employment and tax fraud, etc.,” said Brill. “That’s not to say that credit monitoring is useless because it’s a valuable tool when it aligns with the type of data exposed. Rather, companies will need to gain a better understanding of their actual breach risks, how the breach could actually affect their customers, and the best way to remedy those specific risks and provide better protection to the affected consumers.”

**7. As Cloud and BYOD adoption continues to accelerate, greater accountability will be required for implementing policies and managing technologies.**

The development and evolution of Cloud services and BYOD has moved at a whirlwind pace, leaving IT departments scrambling to get out in front of the technologies and employee usage. In 2014, IT leaders will need to work closely with senior leadership and legal counsel to adapt corporate policies in a way that addresses changing legal risks, while effectively meeting the need of the organization.

“Up until now, cloud and BYOD adoption has been like the Wild West – uncharted, unregulated, and few restrictions. However, we’re seeing courts issue rulings that include significant penalties where discovery, disclosure and other legal obligations aren’t being met because of the use of these technologies,” said Brill. “While it’s implausible to anticipate every possible risk presented by the use of the cloud and BYOD, companies that have integrated these technologies into their corporate policies, IT security, and risk management plans will be much better prepared to fulfill their legal obligations. Organizations must realize that even if they don’t want to deal with this, they’re not going to have much choice.”

Similar are Grayson Milbourne expectations - the Security Intelligence Director for Internet security company Webroot, also Jim Ivers expectations - Chief Security Strategist, Covata and Charles McColgan, Chief Technology Officer of TeleSign.

**Conclusion**

Technological and organizational measures to protect the data in an organization are the basis for building a good IT concept for security. Any company that strives to ensure business continuity, centralized management, resource optimization, quality products and services

ensuring is needed of such a concept. Very often the slightest information loss can lead to adverse financial results and for the IT security consideration on par with all business processes should be a priority for any organization. Protection measures can be directed to the protection of information from a large number of threats in order to ensure business continuity, minimize possible damage and maximize return on investment and business opportunities. Information security is expressed as the sum of physical security, documentary security, industrial security, personal security and security of information systems and networks. Therefore, knowledge of the issues will support for more optimal solutions and will lead to company losses reduction.

### References

1. Verizon 2013 Data Breach Investigations Report. [http://www.verizonenterprise.com/resources/reports/rp\\_data-breach-investigations-report-2013\\_en\\_xg.pdf](http://www.verizonenterprise.com/resources/reports/rp_data-breach-investigations-report-2013_en_xg.pdf)
2. Арнаудов, Д., А. Крумова. Сигурност и защита на информационните системи, Унив. изд. на ВСУ „Черноризец Храбър“, Варна, 2007
3. <http://www.kroll.com/resources/reports/2014-cyber-security-forecast/>
4. <http://www.verizonenterprise.com/pcireport/2014/>
5. [www.networkworld.com/](http://www.networkworld.com/)
6. [http://cio.bg/5963\\_fishing\\_2013\\_pomalak\\_obem\\_potochni\\_celi&ref=cat](http://cio.bg/5963_fishing_2013_pomalak_obem_potochni_celi&ref=cat)
7. [cio.bg/security/1](http://cio.bg/security/1)
8. [mashable.com/category/heartbleed-bug/](http://mashable.com/category/heartbleed-bug/)

### Authors:

Assoc. Prof. eng. math. **Mariya Nikolova Bruseva**, PhD  
Varna Free University,  
Tel: +359 52 359 572,  
Mob. +359 897 80 55 02, [mariya.bruseva@gmail.com](mailto:mariya.bruseva@gmail.com)

**Antonina Ivanova Ivanova**, Assist. Prof.  
Varna Free University, [akrumova@abv.bg](mailto:akrumova@abv.bg)

## ELECTRONIC RESEARCH ON JUDICIAL-EXECUTIVE ACTIVITIES

Eddie Chakarov, Kaloyan Iliev, Dimitar Georgiev

**Abstract:** *The paper presents a new generation product, which features the computerization of technological processes instead of merely the automation of judicial activities. A model of electronic communication between one application running in the offices of the judicial officers and private bailiffs and several WEB applications has been designed. The communication is realized by extracting the data from the client-server application and automatically filling them into the WEB forms of the electronic registers, supported by the State administration. The most important benefits of the application are: Effective electronic communication between the national and the local registers of various state, judicial and public institutions; Structuring and automating the data transfer and as a result, reducing the number of the unintentional technical mistakes of the users; Acceleration of the information exchange between institutions; Increase of the efficiency of public and private enforcement. The results of the applications are saved back to the client server application databases.*

**Keywords:** *electronic communication, web application, judicial officers, JES*

*The Electronic Government is based on efficient, accurate and reliable communication between the state, the judicial and the public institutions and the citizens.*

*The simplest type of such communication is the exchange of secure and reliable electronic messages signed with some type of electronic signature. Examples of such communication are the VAT declarations submitted each month to the NRA by all the entities and individuals registered for VAT in the country. Despite the obvious advantages, one disadvantage of this type of communication is the duplication of inputting the same data by various components in the process.*

*A second way of electronic communication is the data exchange between end applications running in different structures. A relevant example is connecting the information systems of the Customs and the Revenue Agency. As we know, building such "connectivity" is an extremely expensive and time consuming process in which some technical but predominantly organizational problems should be resolved and dealt with.*

*A third way is for various information systems to communicate via web services by using a variety of communication protocols for automated data exchange. This is a very reliable and effective means for data exchange /see e.g. [1]/. Its disadvantage, as in the previous method, is the necessity of synchronized application development on both sides of the communication.*

*Recently, more and more governmental and judicial structures use other means of communication with citizens and businesses, namely*

remote, authorized periodic WEB form completion. The goal and the great benefit for the State administration of such applications is exporting the data of the respective replenishment DB back to the source that creates them. Data security and verification of their source is achieved by signing the forms with some kind of electronic signature, most frequently with a universal electronic signature/UES/. The downside, as with the first type of communication, is the double manual filling of information both in the WEB form and in the local application that may exist in the Office of a legal entity. This duplication can lead to unintentional errors in the entered data and can also lead to discrepancies in the two databases – the local and the global. Furthermore, a lot of time is wasted in entering the same information twice or more. Such duplication has been avoided, for instance, as demonstrated in example [2]:.

The paper presents a model of communication with different type of applications. In some of them the information is retrieved from a database maintained by a client-server application and is automatically filled into the WEB form issued by the corresponding application. The client-server application is the software system “JES” which automates record-keeping activities in the offices of state and private enforcement.

Several programming problems related to the completion of the WEB application forms have been successfully resolved.

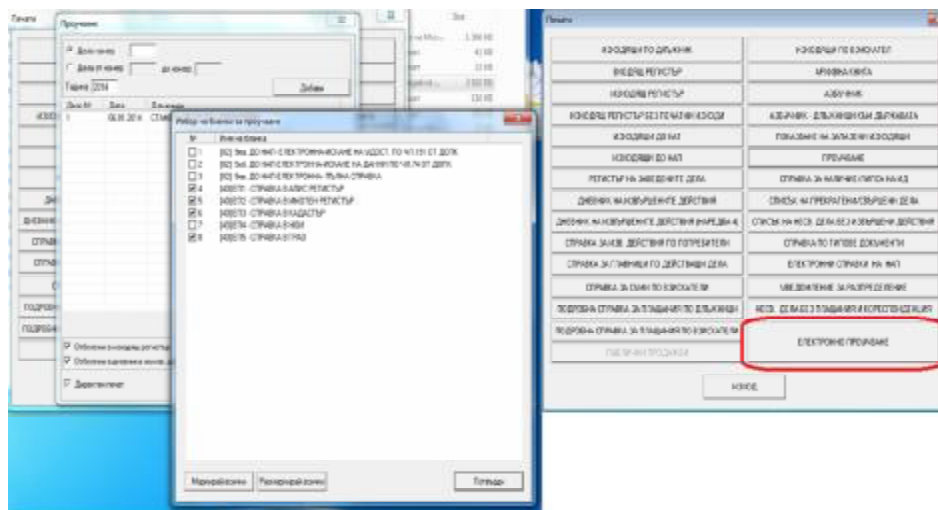
Creating various events and recording data in different destination types in the web form is automated by the software. The first destination type is text fields, whose completion is the most obvious. Other field types include check boxes, hidden fields, radio buttons, dropdown menus and more. A further difficulty is that the forms submitted to some of the applications must be signed with different types of certificates. Some of the forms must be signed with a universal electronic signature. Another type of certificate is necessary for others, and a third type requires user name and password. Of course, a combination of these methods is also possible.

All these problems are successfully dealt with by creating the application – a specific type of browser that functions simultaneously as a standard browser. It has access to the data of the client-server application and utilizes the knowledge of the data structure of the software system “JES”.

This is the browser "SmartBot. It is an application for automatic filling of certain WEB forms with locally applied data. The browser requires that Microsoft .NET Framework 4.0 is installed,(but not necessarily used), as well as Internet Explorer 8 or a higher version. An encrypted connection is created with a WEB form which is to be completed. The application performs the basic functions of the WEB Browser- it displays the WEB form with all its elements and navigation. It

Another integration into the main application are the unified queries in the database of JES They perform three functions. The first is to complement the search with the information available in their database. The second is to form the data files required to run the browser SmartBot. The third is to process the results and store them in the database. In this way, the output data and the results are subsequently transferred into the electronic folder of JES.

## Preparation of Electronic Research in “JES”





Input of an antirobotic codes



Automatic execution of all references - display intermediate results without stopping



Recording the results in directory of the „JES Folder“



2613215012748Notes.txt - Notepad

File Edit Format View Help

1-Справка „други дела в кантората“ /15.12.2013 01:28:21/ за ЕГН:5409028779 РУМИ - има данни за 5 броя висящи изпълнителни  
1-Справка „други дела в кантората“ /15.12.2013 01:28:21/ за ЕГН:4901028779 СОНИЯ ТИЦОВА ЦОНЕВА- няма други висящи и св  
1-Справка „други дела в кантората“ /15.12.2013 01:28:21/ за ЕГН:510221 КРАСИНА СТЕФАНОВА АТАНАСОВА- няма други вис  
1-Справка „други дела в кантората“ /15.12.2013 01:28:21/ за ЕГН:580221 КРАСНИН ИГНАТОВ КИРЯКОВ- няма други висящи  
1-Справка „други дела в кантората“ /15.12.2013 01:28:21/ за Булстат:20121212 ИЗИ ЕС ООД- няма други висящи и свършени  
1-Справка „други дела в кантората“ /15.12.2013 01:28:21/ за ЕГН:81071 КАЛОЯН - има данни за 2 броя висящи изпълн  
1-Справка Търговски Регистър /15.12.2013 01:28:51/ за ЕГН 81071 КАЛОЯН по дело 20138040400008 няма вписвания  
1-Справка АЛПС /15.12.2013 01:28:54/ за ЕГН 5409028779 РУМИ по дело 20138040400001 няма данни  
1-Справка Имотен Регистър /15.12.2013 01:28:57/ за Булстат 20121212 ИЗИ ЕС ООД по дело 20138040400005 няма вписван  
1-Справка Имотен Регистър /15.12.2013 01:28:57/ за Булстат 20121212 ИЗИ ЕС ООД по дело 20138040400005 няма вписван  
1-Справка Кадастръ /15.12.2013 01:28:58/ за Булстат 20121212 ИЗИ ЕС ООД по дело 20138040400005 няма вписвания

## References

- 38

2. Eddie Chakarov, Ivo Panov Electronic Communication with the Public Administration, Информатиката в научното познание 2012, 27–29 юни, ВСУ „Черноризец Храбър“, Варна

**Authors:**

**Eddie Chakarov**, PhD, Associate Professor, Computer Studies Department, Varna Free University “Chernorizets Hrabar”, +359 886567696, [eddy@abv.bg](mailto:eddy@abv.bg)

**Kaloyan Iliev**, MSc in Computer Science, VMWare, +359888265813, [kaloyan.iliev@abv.bg](mailto:kaloyan.iliev@abv.bg)

**Dimitar Georgiev**, Master of Software Engineering, Source Realms Ltd, +359 886138513, [dimitar@sourcerealm.net](mailto:dimitar@sourcerealm.net)

## A PROTOTYPE OF BUSINESS RULES EXTRACTION FRAMEWORK

Delyan Dimitrov, Dimitar Ivanov, Haralambi Haralambiev,  
Krassimir Manev

**Abstract:** *The knowledge for a given business context is usually encapsulated in the software systems used for this business. Aiming at re-engineering of legacy systems, at a later stage some valuable business information can be extracted only with the help of experienced software developers and business analysts. This process can be significantly reduced by automating the process of extraction of business logic. The first step of this process is the extraction of the business rules. The present paper describes some challenges (both conceptual and technological) during the development of a prototype of business rules extraction framework and our way to meet them.*

**Keywords:** *Business Rules Extraction, Re-engineering of Legacy Systems*

### I. INTRODUCTION

Each software system nowadays can be considered as a valuable source of information for its target business, because of the specific business knowledge encapsulated in it. This knowledge is used to limit all users of the system to identical behavior and therefore - make sure they are all guided by the same concepts [6]. With the dynamical changes of the business environment, the target business area varies and after a period of time the composition of the small changes results in the development of significant volume of source code and documentation, aiming to fulfil the newly arisen requirements. Since programmers tend to have confidence only in the source code, usually they do not update the documentation. Respectively, the documentation is frequently left out of date and remains unusable [2]. The outdated documentation leaves the business people without a clear description of the rules, incorporated in the system. However, the system's source code encapsulates the main business knowledge for the target business at its current state and can be useful and reliable source.

The maintenance of a system, implemented with legacy technologies, consumes a lot of resources and if there is no documented description of the system workflow, the problem can be considered of a bigger scale. Moreover, the creation of some new business rules can be achieved by modifying the existing business rules. The latter requires understandable description of the existing business rules. In this context, the manual recovery of business rules can face a lot of obstacles [6], especially when talking about a big heterogeneous system, implemented through a specific technology. Despite maintaining it, another way to deal with a legacy software system is re-implementing it using a given technology. This approach is less resource-consuming, if the business rules used of the old system are presented in a clear form and are easy

to reuse. Otherwise, the new version of a working system would require recreation of business rules, which are already built in the legacy code. If an adequate extraction mechanism is available, this effort can be minimized to a higher extent.

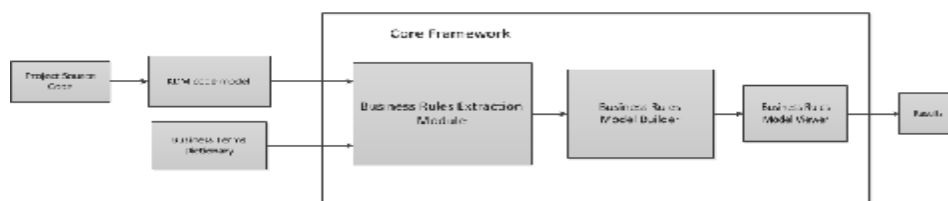
The maintenance and the re-engineering are not the only reason the business rules might need to be extracted from the source code. If the analysis of the business rules built in a system has to be performed by non-technical personnel, it could only be done if a system of business rules that uses domain for the business terminology is present. Key personnel, such as managers, may need to be aware of the processes in the supporting software system, so that they can perform optimizations on the business rules that are applied in it.

The need of an automated cross-language approach is incontestable. This problem is considered by many authors as insolvable, but there are some approaches to solve the task partially and to produce some satisfying results [2]. Even when not solving the problem entirely, it would be extremely difficult to trace business rules without some form of abstractions or decomposition. Fortunately, the model of a business rules system presented by the Business Rules Group (BRG) [7] provides both. The approach described in [4], combined with the business rules classification made by the BRG, is the base for shaping an approach for business rules extraction, later applied in the created prototype system serving this purpose.

## **II. GENERAL APPROACH DESCRIPTION**

The presented approaches for business rules extraction will follow the basic ideas, presented in [4]. The described approach is practically oriented and therefore easy to be applied in a real environment. Moreover, there are clear constraints put on the software system to be analyzed. These constraints delimit a simpler task, part of the more complex task of general business rules extraction from legacy code.

From an algorithmic point of view, most of the challenges of business rules extraction from legacy code do not have optimal solutions. Usually, they can only be solved to some degree, using heuristics, genetic algorithms and non-deterministic algorithmic schemas such as artificial intelligence and neural networks. The purpose of the approach, presented in this paper, is to provide sample implementations of algorithms for most of the problematic areas and to demonstrate how they interact with each other.



**Figure 1. Overview of the framework structure**

Figure 1 describes a sample scheme for integration of the different algorithmic implementations. It also provides a convenient level of decomposition for a general business rules extraction system. From conceptual point of view all modules work independently from each other, so that each can be further modified and improved.

One of the main objectives of the presented prototype is to reduce the human interaction in the process to a minimum. The following sections contain a brief description of the business rules extraction stages and the role of each module, presented in Figure 1. All of the properties, described for a given module apply to its implementation in the prototype. The described sample algorithms below are the ones implemented in it.

#### **A. Business related keywords dictionary**

As it is described above, few of the extraction steps require manual interventions. To clarify the business context, the population of the domain specific dictionary of business terms is important part of the initial process. On a later stage, the extraction of the source code logic depends heavily on the domain-specific business policies, presented in the business dictionary. Even the slightest change in the business terms can change the context and lead to entirely different results. For the extraction process to be independent of this context, an automated approach for retrieving the domain-specific logic from the project can be applied. However, this will not guarantee a consistency of the final results, unless the implemented algorithms for extraction of specific business-related terms from source code are all deterministic.

#### **B. Grouping and transforming code statements to business rules**

The task of grouping code statements and forming business rules is handled by the *Business Rules Extraction Module* (see Figure 1). It produces the main content of the extracted business rules, which is later

used by the succeeding components. This module takes input information for the examined software system, structured in *Knowledge Discovery Model way* [5].

The extraction process begins with static analysis of the code, performed to identify the domain types in compliance with the defined dictionary of business terms. To do so, provided with all user-defined types from the KDM representation of the code, the prototype examines whether terms from the dictionary are contained in their names. After domain types are identified, business rules are extracted for each such type and its connections with other domain types [2]. Next, the dynamic analysis of the code begins with extraction of the entry points [2]. At that moment of our research we did not find an effective and universal method for automated extraction of the entry points. Nevertheless, we suggest using a list of manually defined entry points for the examined software. After such entry points are present, they are examined recursively and a custom control flow graph (CFG) is built for every called function. The custom CFG is constructed from the simple CFG as described in [4].

Let us describe in detail the custom CFG building process. We define a non-branched section as a route in the CFG with maximal length, consisted of only non-branching vertices. In the first step, non-branched sections of the CFG are united into one vertex of the new custom CFG – an elementary derivation [5]. When an edge enters in a non-branched section of the CFG, the section is separated into two vertices, as assuming that the edge enters a single node. Every branching vertex is examined as a separate vertex in the result custom CFG – an elementary constraint [4]. Actually, the resulting custom CFG is a simplified CFG, and therefore the edges between the vertices follow the flow of the program.

After building the custom CFG, the business rules are extracted from it. Here our approach defines that only domain specific elementary constraints and derivations are examined, meaning that their associated code statements contain domain variables, or are anyhow connected to the business. From every elementary constraint a number of action assertions are extracted [1]. The action assertions consist of a condition and actions following its evaluation to a certain value [7]. Ultimately, the number of action assertions extracted from an elementary constraint depends on the number of the values, significant for the logic. These values refer to the condition, which can be evaluated from the action assertion as different actions responding to its different states.

Extracted elementary derivations are separated into parts representing a single business specific action. In the case of the proposed prototype, for each elementary derivation a single action is

aggregated, but several improvements to this aspect of the process would have great impact on quality of the produced results. The business rules formed at this point are passed to the *Business Rules Model Builder*.

#### **C. Combine the extracted business rules with a business model**

The task of the *Business Rules Model Builder* (see Figure 1) is to take the unstructured set of business rules and create an organized hierarchy that represents the business rules model of the underlying software system. A simple sorting of the business rules (by their type) is adequate enough for the purpose of the prototype.

#### **D. Represent the business model**

Visualizing the business rules model is a task that has to be handled differently depending on the expected results. If the rules will be visualized as class diagrams, the representation component will produce such results. The representation task is handled by the *Business Rules Model Viewer* (see Figure 1). It creates understandable visualization of the extracted business rules model. The prototype uses a text representation builder that applies different algorithms to visualize each kind of extracted business rule. The task of creating grammatically correct sentences that represent the extracted business rules without modification is beyond the scope of the presented prototype; therefore the algorithms used to create representations were sample - visualizing only direct properties of the extracted business rules.

### **III. SPECIFICATIONS USED**

To formalize the product of the prototype, we have selected several specifications, which are used as a basis for the representation of the results. The following sections summarize the used specifications and their specifics.

#### **A. Knowledge Discovery Meta-model specification**

In order for the approaches to be cross-language, a generic representation for software systems is needed. The KDM specification provides a common intermediate representation for an existing software system [5]. It describes a meta-model of the key aspects of knowledge, related to the various facets of the enterprise software. The substantial fact here is the practical creation of the code model, described in [8].

#### **B. Business Rules Group specification**

Business rules can often meet various constraints - legal, marketing, technological, etc. Without representation, abstract enough



that decomposes business rules to atomic parts, they might be impossible to be traced. Furthermore, without a clear definition of the important aspects of the target business, the extracted rules will be inconsistent, incomplete and unclear. The BRG specification [7] provides clear description for what constitutes a business rule and distinguishes the basic kinds of business rules. According to it, business rules are divided into three main groups - *structural assertions*, *action assertions* and *derivations* [7]. *Structural assertions* are rules that can be graphically represented and easily formalized, while *action assertions* are constraints on the behavior of the business, which are harder to describe in a formal manner. Thus, the modelling of the action assertions is a significant challenge. *Derivations* are used to describe how present knowledge is used to gain new knowledge [7].

The following section describes briefly the BRG specification, its usage in the prototype and the obstacles, identified in the process of its integration. According to the BRG, a statement of a business rule falls into one of the four:

- **Business term** – used to describe how people think and talk about things. The simplest business rules;

- **Fact** – facts relate terms to each other. The structure of an operating organization can be described using facts that relate terms to each other;

- **Constraint** – also called action assertions. They refer to business rule constraints [1]. The business rules constraints are the rules that constrain how data may or may not be modified. Action assertions take care of checking the integrity of data in a business, or prevent actions from taking place;

- **Derivation** – describe how knowledge from one form is being transformed into other knowledge, possibly in a different form.

Facts consist of two or more terms that have a specific order and each has a fixed role. The fact types described in the specification are:

- **Generalization fact** – used to show when one type defines a subset of occurrences of another type;

- **Attribute fact** – expresses a “has property of” relationship;

- **Participation fact** – describes any kind of relationship between terms other than generalization and attribute. There are three sub kinds of participation facts:

- **Role fact** – used to describe that one type can have specific meaning for another type;

- **Aggregation fact** – the terms in this fact define the component parts of a whole;

- **Association fact** – any participation fact that is not a role or aggregation.

While the structural assertions describe possibilities, action assertions impose constraints. Action assertions consist of an *anchor object* and a *correspondent object* [7]. The *anchor object* can be any kind of business rule, and the *correspondent object* is a *construct*. A *construct* is defined as either a business rule or an action. A simpler explanation of this structure is that in a statement “if...then...” the anchor object would be the phrase that goes after “if”, and the correspondent object would be the phrase after “then”.

There are many criteria for classification of the action assertions. For the means of reverse engineering of facts, we will focus on the approach for splitting the action assertions into integrity constraints, conditions and authorizations [1]. *Integrity constraints* are action assertions that by definition must be true. *Conditions* are business rules that fire other business rules in case a certain condition is satisfied. *Authorizations* are rules of the type “only x may do y”.

In a system of business rules it is important to know how new knowledge is created from the present one. Derivations connect simple facts, forming derived facts. In the BRG specification, the derivations are classified as inferences and mathematical calculations, depending on how the new knowledge is formed. The only difference between the derived facts and all other kinds of facts is that the derived facts contain a derivation. When examining a derived fact, the derivation should not be visible, and so it is not possible to tell apart a derived fact from a simple one.

The described types of business rule statements proved to be adequate in many aspects for the purpose of business rules extraction from source code. Even though one of the initial objectives of the BRG has been for the specification to be used in reverse engineering of business rules, they state that there is much work to be done for this to be fully achieved.

#### a) Usage of the BRG specification in the prototype

For the means of reverse engineering of business rules, it was convenient to split the terms, contained in a fact, into a *dependent term* and a number of *context terms*, with fixed ordering. Their roles are determined by the type of the fact, which connects them. Facts of different types describe different relationships in the code. *Generalization facts* are used to identify inheritance amongst types of significance. If a domain type extends another domain type [2], it defines a subset of its occurrences. Relationships showing that some domain type has an attribute of another domain type are represented by *attribute facts*.

*Association facts* are used to describe variety of relationships between domain types, such as function invocation associated with

parameters, association between a declared variable and the parameters of its constructor, and so on. For the purposes of better understanding, the association fact type is extended by more specific types, describing the concrete association between the terms involved, since that type of fact itself is too general and therefore the subtypes bring more specific relationships.

The usage of derived facts to describe different assignment structures comes naturally, since they are the places where a piece of present knowledge is used to gain new knowledge. An *inference* is recognized when only one operand is present on the right-hand side of the assignment, and otherwise this assignment is a *mathematical calculation*. Even though this description of the derivation kinds may sound simple, there are many complications involved in their practical use for business rules extraction.

Only two of the main action assertion kinds, described in the BRG specification proved to be product of reverse engineering in the practice – conditions and integrity constraints [7]. Conditions are used to describe the conditional statements with business related logic in a program. The integrity constraints are events that cause complete malfunctioning of a software system, e.g. exceptions. At this stage, we found no suitable mechanism to identify authorizations from a legacy system, as it is hard to find a code construction to be recognized as such.

#### **b) Impediments of modelling the BRG specification**

The BRG defines the main objective of their specification – its usage for reverse engineering of business rules from existing systems [7]. The specification is of great use in the practice and especially for the task for analysis of existing software system and extraction of its business rules. However, as the BRG states, there is much work to be done in the direction of full usability of the specification for such practical purposes (as experimented in the presented prototype) [7]. Some of the business rules, as described in it, proved to be too formal. Such formality is hard to be achieved, if not impossible, through reverse engineering from legacy code systems.

One of the biggest impediments of the business rules modeling introduced by the BRG were the action assertions. They proved to be too formal to serve the means of reverse engineering of business rules from an existing software system.

The conditional statements are another met impediment as they are an important part of every programming language. In general, the condition can be any expression that can be evaluated to true or false. In the case of some programming languages, this definition puts almost no limitations on the operands involved in it. This includes no limitations on

the number of domain type variables involved in such expression. In the practical use, we faced the following problem - the anchor object refers to several variables, all concerned with the business logic. Most of the proposed approaches for addressing this problem would have resulted in radical changes in the used structure of action assertions.

The solution was found on a more detailed action assertions description, described by Hay [1]. His work gives more practical and structured meaning to the action assertions. Hay introduces only two types of action assertions, mentioned above as usable – integrity constraints and conditions. A condition is a rule that may be true or false and controls the firing of other business rules, while an integrity constraint is a rule, which by definition must always be true, about an entity type, class, role, or an attribute [1]. Each business rule constraint is composed of one or more business rule constraint elements. Business rule constraint elements do the constraining and consist of an operator and a description. This structure of the business rule constraints fits the purpose of extracting them from conditional statements in the code, and therefore its integration in the prototype was easier. To modulate business rule constraint elements, a recursive structure is used, presenting a tree of the operations, used to evaluate a given condition to its true or false state, including the operands. The used structure clearly describes the condition and all the constraints laid on domain type variables it involves. The structure of action assertion includes some business rule constraint elements and the corresponding actions, depending on the value that the condition is evaluated to.

Another impediment in the process of extraction was to identify an inference derivation from a mathematical calculation [7]. Even though a simple mechanism for this to be done was described above, there are some obstacles, met in the extraction of this kind of business rules. They were mainly connected with the different usage of functions in assignment constructions. Invocation of functions on parameters can easily be classified as an operator performed on them. However, in object oriented languages this can be ambiguous, since methods can be used as getters of some member variables, in the means of encapsulation. This imposes further analysis on the functions invocations in assignments in order to really understand their purpose.

#### **IV. RESULTS**

The prototype managed to achieve the extraction of key business rules, which describe most of the important processes in a software system, supporting a certain aspect of a business. All results presented are a product of examining the Pygmy [3] project with the prototype business rule extractor. The Pygmy project is a webserver meant for

embedding into applications. Therefore, the used term dictionary was composed of words related to the server logic, such as “*http*”, “*server*” and “*view*”.

Here are some of the facts from the static extraction:

● **Attribute facts:**

- o “*http response contains a stream*”;
- o “*message view contains a topic*”;
- o “*server contains a response listener*”.

● **Generalization facts:**

- o “*server is a runnable*”;
- o “*document view is a view decorator*”.

● **Role facts:**

- o “*abstract handler can be handler for server*”;
- o “*directory handler can be handler for server*”;

● **Association facts:**

- o “*http can get translate path using root and url*”;
- o “*view can get forum url using relative url*”;
- o “*document view can get be rendered using request*”.

The extracted business rules described the true state of the code with no modifications. All rules have been classified by their type and excluded technical processes, specific for implementation language of the software. The human interaction in the process was reduced to a minimum. Since representation of the extracted facts was not the main objective of the prototype, this part might be considered not completely finished and should be an object of future work. The most important steps to achieve such results using the described approach are given.

#### **A. Modelling business rules**

Appropriate modeling of business rules is critical for the extraction process. The structures containing business logic have to enable the system to keep all the needed information in them. Furthermore, the modeled rules must not impose any modification on the extracted logic, so that it would reflect the true state of the source code. The modeled rules have to provide all the needed information for the following stages of the extraction.

#### **B. Identifying domain types and variables**

Providing an accurate mechanism for identifying domain types and variables gives the basis for delimiting business related logic from irrelevant technological processes in the system. Furthermore, facts extracted from domain types are the core facts that describe the fundamentals of a business, as they comprise the information about the consistence of key objects and the part they play in the business

process. With no such identification mechanism the extracted rules will be mixed up with the implementation details, which are not concerned with the business logic. Such systems of rules would often be unusable, because of the volume of technologically specific processes involved in each system.

### **C. Identifying business related statements**

When analysis is performed on a software system of a big scale, it is critical that the final business rules are as filtered as possible, presenting the key business processes. This is because the bigger the software system is, the more information about technology specific processes it contains. If such processes are not identified, we risk that unrelated to the business logic will be examined in the extraction process, which can cause severe performance issues of the framework. On the other hand, if some rules that should be examined are still not, the extraction could miss important for the business specifications.

### **D. Extraction techniques**

Surely, the extraction algorithms used in the prototype and presented above are not the only way to accomplish the business rules extraction. Whatever the used approach is, carefully choosing and testing analysis algorithms, which go through parts of the legacy system, is mandatory. Legacy systems can be huge and performance issues can be faced when parts, irrelevant to the business process are examined or when the examination process is too complex and slow. Because of this, only parts of the code have to be examined when this is possible without the loss of important business related information. If recursive functions are product of examination, it has to be done with caution, since they can cause infinite loops if not appropriately approached. Grouping code statements into one whole process, hiding the unimportant details from the end user is also a key part of the extraction that can be approached differently.

### **E. Structuring of the extracted rules**

Rules that are not structured properly may be presented in a sequence that will make them unintelligible, or even in a random order. This is why they should be structured and ordered in a way that will properly serve the purpose of the extraction and will make them easy to read and process. Such ordering may greatly vary with the purpose and requirements of the extraction and needs to be considered individually for each particular case.

#### **F. Representation of the extracted rules**

This is the final part of the process and its result is not a product of further modification. It is responsible for displaying the extracted business information in a way, appropriate for the end user. This representation should suite well to the purpose of the extraction performed. According to the requirements for such representation, the rules can be presented as text, class diagrams and so on. Even if the extracted rules reflect all key aspects of the business, their improper visualization can make the whole process senseless.

#### **V. CONCLUSION**

The presented prototype proved the complexity of the tasks, concerning the business rules extraction from legacy systems. Moreover, it illustrates how the theoretical approaches for business rules extraction can be combined to solve more practical problems and produce basic results with concrete value. The modular structure of the prototype ensures that the future work in the domain will give more efficient results by improving the existing heuristics, algorithms and methods, encapsulated in the different extraction modules. On the other side, the results shown indicate which modules of the prototype will need more improvements in the future.

#### **VI. ACKNOWLEDGEMENT**

This work is supported by the National Scientific Research Fund of Bulgaria under the Contract ДТК 02-69/2009 “Automated extraction of business rules and processes from software code (Business Rules Mining)”.

#### **References**

- [1] Hay, D. C., A Repository Model - Business Rules (Action Assertions), April 1, 2012, retrieved from <http://www.tdan.com/view-articles/4987>, last accessible on 14.03.2012
- [2] Huang, H., W. T. Tsai, S. Bhattacharya, X.P.Chen, Y. Wang, J. Sun, Business Rule Extraction from Legacy Code, in proceedings of the 20th Conference on Computer Software and Applications, p.162, USA, 1996.
- [3] Hubbard, C., Pygmy HTTP Server Project, <http://pygmy-httpd.sourceforge.net>, last accessible on May 01, 2014
- [4] Manev, K., N. Maneva, H. Haralambiev, Extracting Business Rules Through Static Analysis Of The Source Code, in proc. of the Forty First Spring Conference of the Union of Bulgarian Mathematicians Borovetz, Vol. 41, No 1, (2012), 247p-253p April 9–12, 2012.
- [5] OMG, Knowledge Discovery Metamodel (KDM), <http://www.omg.org/spec/KDM/1.0/>, January, 2008

[6] Ramsey, F. V., J. J. Alpigini, A Simple Mathematically Based Framework for Rule Extraction from an Arbitrary Programming Language, in proceedings of the 26th International Computer Software and Applications Conference on Prolonging Software Life: Development and Redevelopment, pp. 763-772, Washington, DC, USA, 2002

[7] The Business Rules Group, Defining Business Rules - What Are They Really?, GUIDE Business Rules Project, Final Report, rev. 1.3., July, 2000.

[8] Yanakiev, A., H. Haralambiev, K. Kraichev, Revisiting Abstract Syntax Tree as a Basis of Source Code Knowledge Models, Informatics in the Scientific Knowledge, pp.242 – 257, Varna, Bulgaria, 2012

**Authors:**

Delyan Georgiev Dimitrov, Applied Research and Development Center, Musala Soft, +359 (888) 826 246, [delyan.dimitrov@musala.com](mailto:delyan.dimitrov@musala.com)

Dimitar Stoyanov Ivanov, Applied Research and Development Center, Musala Soft, +359 (884) 434 452, [dimitar.ivanov@musala.com](mailto:dimitar.ivanov@musala.com)

Haralambi Kostadinov Haralambiev, Applied Research and Development Center, Musala Soft, +359 (889) 440 890, [haralambi.haralambiev@musala.com](mailto:haralambi.haralambiev@musala.com)

Krassimir Nedelchev Manev, Assoc. Professor, Ph.D., Department of Informatics, New Bulgarian University, +359 (887) 253 211, [krmanev@gmail.com](mailto:krmanev@gmail.com)



## SIMULATION MODEL THE DYNAMICS OF THE SYSTEM ARISING FROM INTERACTION OF TOURISM WITH THE LOCAL ENVIRONMENT

Viktors Gopejenko, Yelena Kozmina

**Abstract.** *Tourists are attracted by the beauty of the surrounding World. Corresponding infrastructure is created to service them – hotels, airports, and roads. The more there are tourists, the larger infrastructure is required, however, the more the environment loses in its attractiveness, which leads to a limit of the tourism development. There are two possible scenarios of managerial decision-making in this situation – mild with maximum preservation of the environment's attractiveness and careless that leads to the destructions of the tourism profitability basis.*

*System dynamics computer modeling results that are caused by the interactions of the tourism with the local environment are considered in this paper.*

**Keywords:** *System Dynamics, Logistic Function, Computer Simulation, Vensim.*

Attractive natural places draw tourists' attention. The more there are tourists, the more money is spent for their servicing, the more damage is inflicted to the surrounding eco system, and the more difficult it is to restore its attractiveness. The loss of the attractiveness leads to the decrease of the tourist flow, to the degradation of the area and to the ruined environment.

Which conditions are required so that the area could be uninterruptedly used for tourism without loss of the attractiveness? Computer modeling might help to understand system dynamics, which appears during the interaction between tourism and local environment.

Similar approach to study dynamic systems was considered in refs. [1,2,3].

Let – qualitative characteristics of the environment, taking into account the infrastructure created at time  $t$ ,  $K$  - maximum possible value of the qualitative characteristics of the environment.

Assuming that the change rate of the environment is proportional to the quality of its current state  $q(t)$ , i.e.  $\frac{dq}{dt} = kq$ , and the proportionality coefficient  $k$ , changes proportional to the difference  $K - q$ , i.e.

$$k = p(K - q) = pK \left(1 - \frac{q}{K}\right).$$

Taking into account mentioned above, the equation can be written as follows:

$$\frac{dq}{dt} = r q \left(1 - \frac{q}{K}\right), \quad (1)$$

where  $r = pK$  – proportionality coefficient that describes the growth rate of the quality characteristics of the environment,  $r > 0$ . Initial condition

$$q(0) = q_0.$$

Note that function  $q = 0$  and  $q = K$  at all  $t$  are the solutions of the given equation, that is why the equation is solved for  $q \in (0; K)$ . Making a substitution in the equation

$$\frac{q}{K} = Q, \quad Q \in (0; 1), \quad Q_0 = \frac{q_0}{K}.$$

Getting

$$\frac{dQ}{dt} = rQ(1 - Q).$$

Integration gives the following solution:

$$Q = \frac{Q_0 e^{rt}}{1 + Q_0(e^{rt} - 1)}.$$

Solution behavior at infinity:

$$\lim_{t \rightarrow +\infty} Q(t) = \frac{\lim_{t \rightarrow +\infty} Q_0}{\frac{1}{e^{rt}} + Q_0 \left(1 - \frac{1}{e^{rt}}\right)} = 1.$$

Substituting  $Q = \frac{q}{K}$  and  $Q_0 = \frac{q_0}{K}$  in the solution

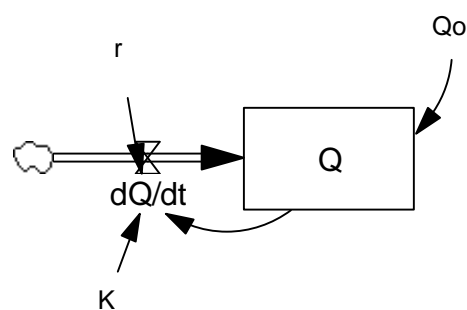
$$Q = \frac{Q_0 e^{rt}}{1 + Q_0(e^{rt} - 1)}, \text{ getting a logistic function:}$$

$$q(t) = \frac{Kq_0 e^{rt}}{K + q_0(e^{rt} - 1)}$$

$$\lim_{t \rightarrow +\infty} q(t) = K.$$

### Model 1. Depends of environment quality on time.

Let's construct a model of Vensim PLE for the differential equation (1), Fig.1



$$\begin{aligned} \text{"dQ/dt"} &= r * Q * (1 - Q/K) \\ K &= 1.5 \\ Q &= \text{INTEG}(\text{"dQ/dt"}, Q_0) \\ Q_0 &= 1 \\ r &= 1 \end{aligned}$$

Fig. 1a. Vensim PLE model for the differential equation (1)

Fig. 1b. Equations of the model Fig.1a.

We have used normalized state of these variables since at first we are interested in behavior of the estimating system.

Without tourism environment regeneration rate  $r$  the environmental quality  $Q$  increases  $\left(\frac{dQ}{dt}\right)$ , and environmental quality is changing as a logistic function limited by environmental carrying capacity  $K$  Fig. 2a, b.

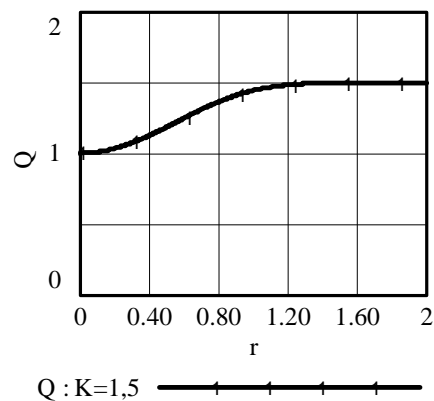
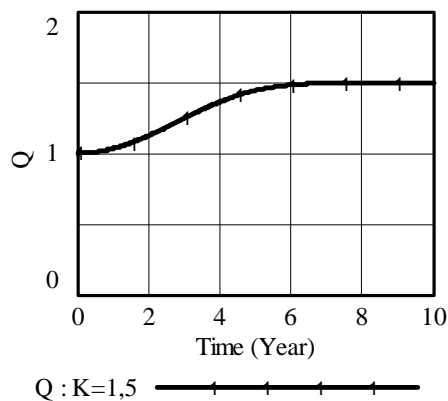


Fig. 2a. Dependences of quality (Q) from time,  $K=1,5$

Fig. 2b. Dependences of quality (Q) from environment regeneration rate (r), phase portrait,  $K=1,5$

## Model 2. Correlation between the quality of the environment and the number of tourists

Let's assume that  $N = N(t)$  – is the number of tourists that use given environment with infrastructure created within it. Suppose that the change rate of the number of tourists increases proportionally to the quality of the created object and decreases proportionally to the number of tourists that have already used this service:

$$\frac{dN}{dt} = aQ - bN,$$

where  $a$  and  $b$  are corresponding proportionality coefficients,  $a, b > 0$ . In the presence of tourists quality of the environment is getting worse in proportion to the number of tourists. Let's assume that proportionality coefficient depends on the quality of the existing quality of the environment. Tourists and environment quality cause environmental stress, which decreases the environment quality according to the specific environmental degradation rate ( $d$ ), then:

$$\frac{dQ}{dt} = rQ(1 - Q) - dQN, d > 0,$$

Getting the system of differential equations:

$$\begin{cases} \frac{dQ}{dt} = rQ(1 - Q) - dQN, \\ \frac{dN}{dt} = aQ - bN, \end{cases} \quad (2)$$

where initial conditions are  $Q(0) = Q_0$ ,  $N(0) = N_0$ .

Let's determine the equilibrium state for environ quality ( $Q$ ) and tourists ( $N$ ) as function of advertisement effect ( $a$ ) and environmental deterioration rate ( $d$ ) analytically and plot the results as a function of these parameters.

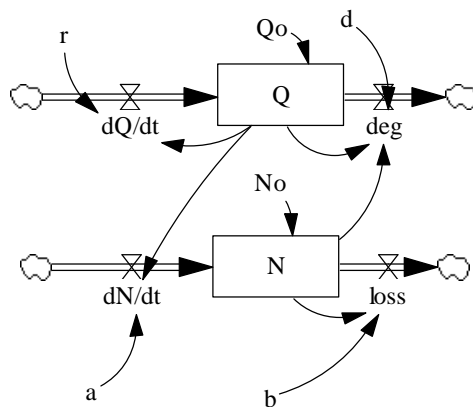
Equilibrium state of this system is found by solving the system of the equations

$$\begin{cases} rQ(1 - Q) - dQN = 0, \\ aQ - bN = 0. \end{cases}$$

Besides the obvious solution  $\begin{cases} Q = 0, \\ N = 0, \end{cases}$  there is also the following solution:

$$\begin{cases} Q = \frac{br}{(br + ad)}, \\ N = \frac{ar}{(br + ad)}. \end{cases}$$

The Vensim PLE model for the system of differential equations (2) is shown on Fig 3.



```

a=      5
b=      0.5
d=      1
deg=    d*N*Q
"dN/dt"=a*Q
"dQ/dt"=r*Q*(1-Q)
loss=   N*b
N= INTEG ("dN/dt"-loss, No)
No=     0.1
Q= INTEG ( "dQ/dt"-deg,
Qo)
Qo=     1
r=      1

```

Fig. 3a. A model of Vensim PLE for the system differential equation (2)

Fig. 3b. Equation of the model Fig.3a.

Fig. 4a presents the plots of environ quality ( $Q$ ), tourists ( $N$ ) for the default parameter setting (Fig. 3b). We can see that the location of the stable point is independent of initial conditions.

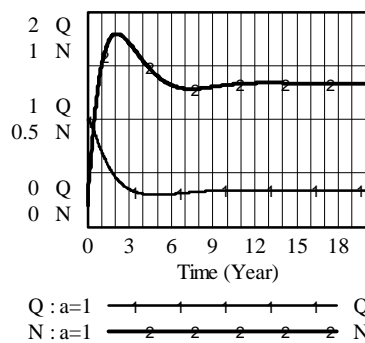


Fig. 4a. Dependences of quality (Q) and tourists (N) from time

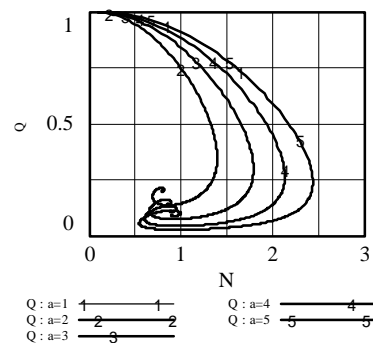


Fig. 4b. Dependences quality (Q) from tourists (N), phase portrait.

A state diagram describing the dependence of the environment quality on tourists shown on Fig. 4b. demonstrates state trajectories and equilibrium points' locations as a function of advertisement effect (a). It is seen that a larger advertisement scale causes the increase of the number of tourists, however, this also causes a significant negative impact on the quality of the environment.

Fig. 5a. clearly shows that system development's dynamics and its equilibrium state in particular heavily depend on the advertisement effect. The evolution of  $Q$  and  $N$  shown in Fig. 5b. is significantly affected by the worsening rate of the environment caused by tourism. These parameters are also heavily dependent on the environment regeneration rate (d), which defines the number of tourists that can be persistently sustained by the environment.

The time interval during which the number of tourists increases is also of interest. It is necessary to take additional administrative decision at the moment when the number of tourists starts decreasing.

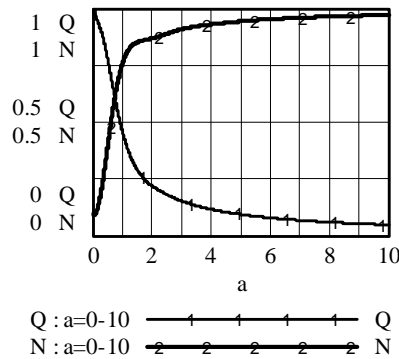


Fig. 5a. Dependences of Quality (Q) and tourists (N) on advertisement effect (a), phase portrait.

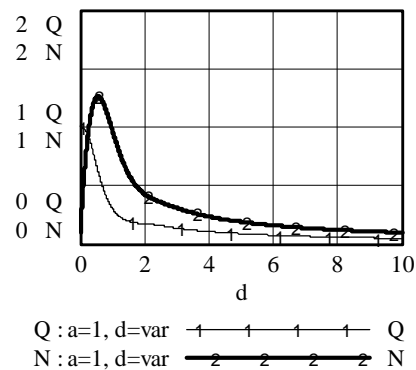


Fig. 5b. Dependences of Quality (Q), tourists (N) on degradation effect (d), phase portrait.

### Model 3. A correlation between the quality of the environment, the number of tourists, and financial investments.

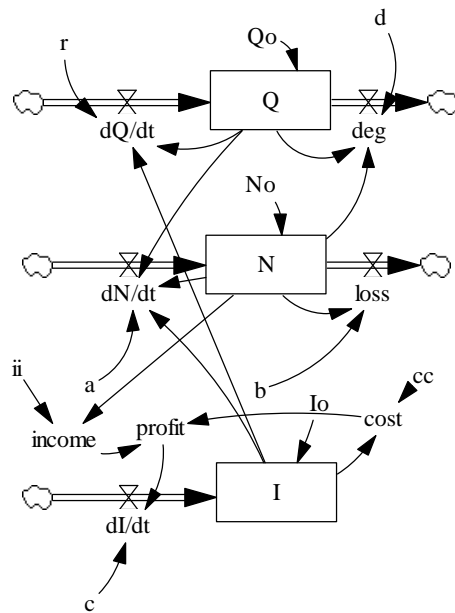
Let's add one more function to consider  $I = I(t)$  – financial investments that are used in this process. Let's assume that the change rate of  $I(t)$  function increases proportional to the number of tourists and decreases proportional to the financial means involved in the turnover, i.e.  $dI / dt = cN - pI$ . Let's substitute advertisement effect  $a$  and environ-

ment regeneration rate  $r$  with  $a/I$  and  $r/I$  correspondingly. This results in the following system:

$$\begin{cases} \frac{dQ}{dt} = rIQ(I-Q) - dQN, \\ \frac{dN}{dt} = aINQ - bN, \\ \frac{dI}{dt} = cN - pI, \end{cases}$$

with initial conditions  $Q(0) = Q_0$ ,  $N(0) = N_0$ ,  $I(0) = I_0$ .

Model Vensim PLE for system (3) is shown on Fig. 5. This model was expanded so that the annual profit ( $income - costs$ ) can be calculated.



```

a= 9
b= 0.4
c= 0.02
cc= 0.2
cost= I*cc
d= 0.1
deg= d*N*Q
"dl/dt"=c*profit
"dN/dt"=a*Q*I*N
"dQ/dt"=r*Q*(1-Q)*I
I= INTEG ( "dl/dt", lo)
ii= 0.6
income=N*ii
lo= 0.1
loss= N*b
N= INTEG ("dN/dt"-loss, No)
No= 0.1
profit= income-cost
Q= INTEG ("dQ/dt"-deg, Qo)
Qo= 1
r= 0.1

```

Fig. 5a. Vensim PLE model for the system of the differential equations (3)

Fig. 5b. Equation of the model Fig. 5a.

Fig. 6 presents the plots of simulation for the environ quality (Q), tourists (N) and investment (I) and the total profit for the default parameter setting.

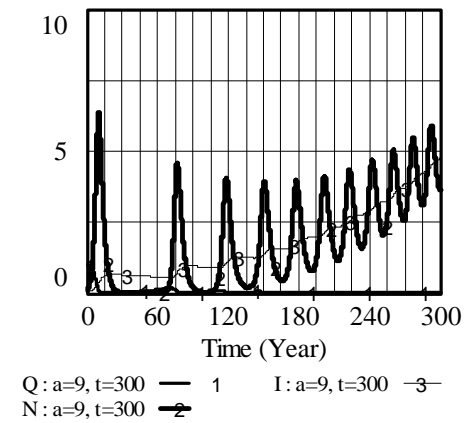


Fig. 6a. Dependences of Quality (Q), tourists (N) and investment (I) on Time

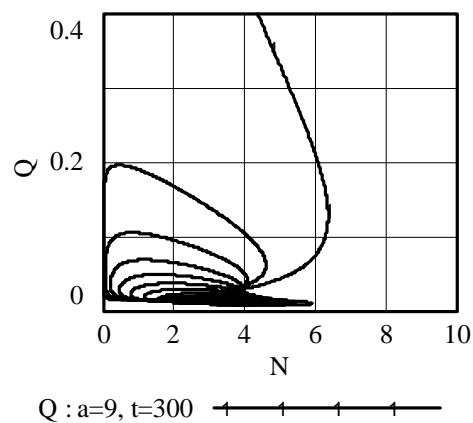


Fig. 6b. Dependences of Quality (Q) on tourists (N), phase portrait

Note that correlation of “quality – number of tourists” Fig.6 a. fluctuates near the curve that characterizes financial investments.

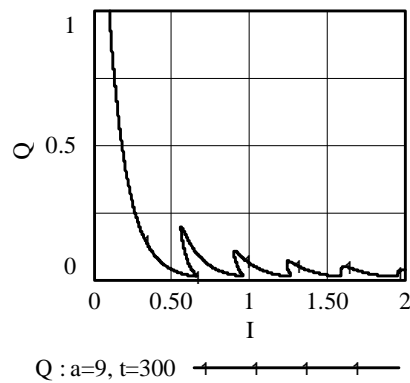


Fig. 6c. Dependences of Quality (Q) on investment (I), phase portrait

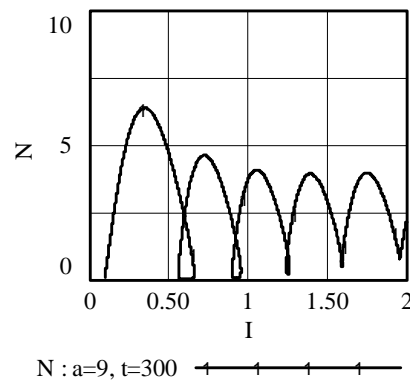
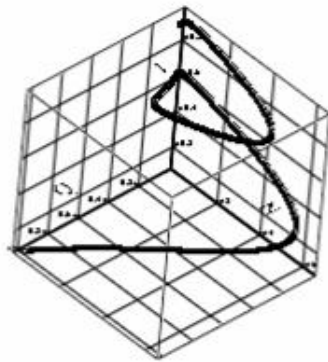


Fig. 6d. Dependences of tourist (N) on investment (I), phase portrait





(Q, N, I)

Fig. 6e. 3D plot dependences of Quality (Q), tourists (N) and investment (I), phase portrait

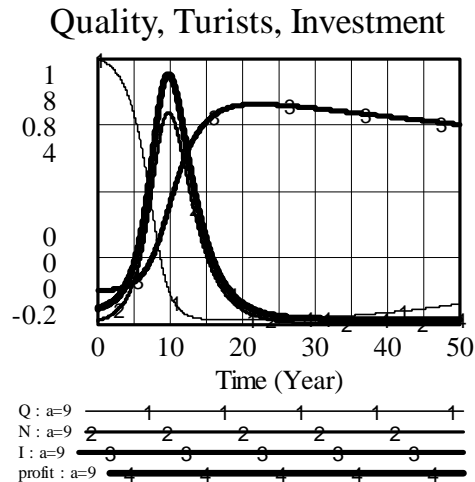


Fig. 6f. Dependences of Quality (Q), tourists (N), investment (I) and profit on Time

## Conclusion

We can examine the behaviour of the model above as a function of their parameters, identify those parameters that have little influence on the behaviour, concentrate further investigations on the parameters that strongly influence the behaviour, and document prominent behavioural mode (oscillations, collapse, equilibrium states) as a functions of these parameters.

The computer simulation based on Vensim allows to develop an advertising and investment strategy, which produces maximum total profit and to get visualizing solutions of the differential equations dynamically.

## References

1. Hartmut Bossel (2007). System Zoo 2 Simulation Models. Climate, Ecosystems, Resources: Norderstedt, Germany, pp.192.
2. V. Gopejenko, Y. Kozmina (2012) System Dynamics and System Thinking in Educational Process. The 10th International Conference Information Technologies and Management 2012. April 12-13, 2012, Information Systems Management Institute, Riga, Latvia, p.51-52.

3. V. Gopejenko, Y. Kozmina. Computer Simulation and System Dynamics in Educational Process (2012). International scientific conference: INFORMATICS IN THE SCIENTIFIC KNOWLEDGE June 27-29, 2012, Varna Free University, Varna, Bulgaria, pp. 305-320.

**Authors:**

Professor, Dr. sc. Ing. Viktors Gopejenko,  
As. Professor, Dr. Math. Yelena Kozmina  
Information Systems Management Institute  
Department of Natural Sciences and Computer Technologies  
Lomonosova 1 k7, LV-1019, Riga, Latvia,  
e-mail: [viktors.gopejenko@isma.lv](mailto:viktors.gopejenko@isma.lv); [jelena.kozmina@gmail.com](mailto:jelena.kozmina@gmail.com)

## PROTECTION OF PATIENT'S RIGHTS IN eHEALTH ENVIRONMENT

Iskra Mircheva

**Abstract:** *Preserving the privacy and confidentiality of the medical data of the patients has always been a fundamental issue in medicine and healthcare. This issue remains open and much more important with the implementation of eHealth. This paper outlines some basic issues of preserving the patient's right for privacy and confidentiality of his or her medical data originating from the challenges of the growing implementation of information technologies in medicine and healthcare. Some confidentiality and security measures as well as the basic requirements to the security of electronic medical data are discussed.*

**Keywords:** *patient's rights, privacy, confidentiality, information technologies, electronic medical record, eHealth*

### Introduction

The electronic storage of vast quantities of medical data and their organization in electronic medical records, the opportunity to communicate medical data among the different healthcare providers through specialized and public computer networks, the development of different medical information systems – these are only a small part of the implementation of information technologies in medicine and healthcare.

While supporting professionals in the delivery of healthcare, eHealth information systems also have the potential to empower the patient. However, to the extent that contractual relations which can be influenced by the patient replace state supervision and responsibility, new requirements for the protection of the individual's rights emerge.

Preserving the privacy and confidentiality of medical data is an issue that for centuries has been considered fundamental in medicine and healthcare. Ever since the 4<sup>th</sup> century BC the medical professionals take the oath of Hippocrates that obliges them to keep in secret the information they have obtained from their patients during the process of delivering health care.

The ever-growing implementation of information technologies in healthcare and medicine, especially the development of the electronic medical record and the connection of clinical databases as well as the freedom these systems offer in relation to their usage force the development of new policy and security procedures for protection the privacy and confidentiality of medical data. The increased use of information technologies and computer applications in medicine and healthcare makes the issue of data protection and security even more important especially when such applications concern patient care and personal patient medical data. Besides, the new technologies offer new opportunities as well as new threats that require new confidentiality and security measures.

No matter how far the technologies go, the problem of preserving the privacy and confidentiality of the medical data persists. It has already become evident that the major problem of information security in healthcare is not the technology but the lack of flexible data security policy. It determines the scope of data to be protected, the level of data protection to require and the access to the different types of medical data. The role of the policy is to balance the functional requirements for the system security, i.e. to balance security and confidentiality measures and this is very difficult to achieve.

**The aim** of this paper is to analyze the confidentiality and security measures in order to preserve the patient's rights for privacy.

### **Materials and methods**

Different measures for confidentiality and security are analyzed. The confidentiality measures are in concordance with the patient's right for privacy of his or her medical data and are connected with the legal issues and regulations. The security measures concern the healthcare system and the responsibility of the organizations and the different functional requirements. These are the practical steps to be fulfilled.

Some results are presented based on enquiry carried out among the physicians and nurses in Bulgaria including questions on access, security and legal issues for medical data. Respondents of the study were the physicians and the nurses employed in 5 university and regional hospitals in North-East Bulgaria in 2011. The number of the participants in the study was 650.

### **Discussion**

Preserving the privacy and confidentiality of the medical data of the patients has always been a fundamental issue in medicine and healthcare. This issue remains open and much more important with the implementation of eEurope 2005 Action Plan - An information society for all, followed by the i2010 Digital Agenda for Europe. The Digital Agenda for Europe is one of the seven flagship initiatives of the Europe 2020 Strategy, set out to define the key enabling role that the use of information and communication technologies will have to play if Europe wants to succeed in its ambitions for 2020 [10].

However, the computerization of medical data and information and especially the implementation of computer networks in medicine and healthcare cannot be achieved if the individuals are not convinced that there exist sufficient and adequate measures for the protection and the security of the privacy of their personal medical data. Some of the aspects of the design and implementation of eHealth information systems as pointed out in the Patients Charter for eHealth [1] are access to

appropriate information, security and safety of information systems, mastery of the medical record, privacy and confidentiality, dignity and respect, products from the internet and world wide web, comments and complaints and compensation.

**Confidentiality measures.** The basis of the security of electronically preserved medical data is the recognition of patient's rights for privacy and confidentiality of his or her personal medical data. The main issue is that recognizing the patient's rights for privacy we can understand the tendencies of the confidentiality measures. Therefore, the confidentiality measures are the response to the proper understanding of the patient's rights for privacy. In order to achieve a balance in confidentiality we can point out three different activities.

**The first** activity concerns the development of universal patient identifier. The shared use of medical data by different healthcare providers through computer networks brings forward the importance of developing a universal patient identifier for the whole healthcare system. Although this is considered a major problem in most of the countries, in Bulgaria it has been partially solved. All Bulgarian citizens and permanently residing foreigners have a 10-digit unique personal identification number (UIN), consisting of the date of birth, the gender and the place of birth (the region in Bulgaria). The last digit is for testing the correctness of the UIN. This UIN is a convenient instrument for organization of different sets of personal information, for registry, statistics etc. Although this UIN is not absolutely safe, it is a convenient means for unique identification of all Bulgarian citizens. This identification number is used in all personal documents and records including the healthcare and health insurance systems. However, the use of this UIN as a unique patient identifier is not appropriate. First, it is a number, and it is universally known that every number used as personal identifier can easily cross the boundaries of its initial use and spread over to other domains. On the other hand, it is written along with the names of the individual on a number of personal identification documents (passport, driving license etc.). For the purposes of preserving the privacy and confidentiality of the individual's medical data, the UIN has to be extended with some additional personal patient information thus making it more secure.

**The second** activity is the development of standards for confidentiality measures. This concerns the standards for medical data privacy and confidentiality, as well as the standards for access to electronically stored, processed and transferred medical data. Such standards have already been developed in the USA. Some of the oldest standards are E31.17 [2] - Privacy, access and confidentiality, E31.20 – Security of data and health information systems, E31-22 – Medical transportation and documentation. A number of HIPAA Security stan-

dards have also been developed. They require physicians to protect the security of patients' electronic medical information through the use of procedures and mechanisms that protect the confidentiality, integrity, and availability of information. As of 2005, physicians must have in place administrative, physical, and technical safeguards that will protect electronic health information that the physician collects, maintains, uses, and transmits. The new HIPAA rules are effective since September 23, 2013.

In Europe, CEN/TC 251 (Medical informatics), focusses on the adaptation, profiling and adoption of interoperability standards in the domain of health informatics. CEN/TC 251 engages with most of the standards development organizations in the field of medical informatics. It is a founding member of the Joint Initiative Council alongside HL7 and ISO. This important Council has now expanded to include IHTSDO (develops terminology), CDISC (deals with clinical research) and GS1 (deals with identification). It is currently working with the Initiative by healthcare professionals and industry to improve the way computer systems in healthcare share information (IHE) in project work and with new consortia such as the Clinical information modeling initiative (CIMI). The combined efforts of these organizations are expected to promote the coordinated use of established standards such as DICOM and HL7 to address specific clinical needs in support of optimal patient care and deliver the set of standards that are needed for true interoperability of health information systems.

CEN/TC 251, Work Group 3 – Security, safety and quality, deals with the development of techniques for technical protection of confidentiality, integrity, availability and accountability as well as guidelines for security management of medical data. A number of such standards have been approved: CR 13694:1999 Health Informatics - Safety and security related software quality standards for healthcare (SSQS), CR 14301:2002 - Health informatics - Framework for security protection of healthcare communication, EN 12251:2004 - Health informatics - Secure user identification for health care - management and security of authentication by passwords, EN 13606-4:2007 Health informatics - Electronic health record communication - Part 4: Security (this part of the 6-part standard on electronic health record communication describes a methodology for specifying the privileges necessary to access electronic health record data). Other standards include CEN/TR 15640:2007 - Health informatics - Measures for ensuring the patient safety of health software, CEN ISO/TS 14265:2013 Health Informatics - Classification of purposes for processing personal health information (ISO/TS 14265:2011) - defines a set of high-level categories of purposes for which personal health information can be processed, EN

ISO 27789:2013 Health informatics - Audit trails for electronic health records (ISO 27789:2013), specifies a common framework for audit trails for electronic health records, in terms of audit trigger events and audit data, to keep the complete set of personal health information auditable across information systems and domains.

Such standards in Bulgaria are not yet developed. Some of the standards of CEN/TC 251 are acknowledged but actually they are not applied.

**The third** activity concerns the legal issues. The general aim of the legal issues is to (1) define the rights of the individuals concerning their health and medical data, created or maintained as a part of the treatment process; (2) to define the rights and the responsibilities of the individual who creates or maintains individually identified health or medical information created during a treatment process; and (3) to introduce effective mechanisms for fulfillment of points (1) and (2).

The Confidentiality Act in USA [15] deals with the privacy of the individuals concerning his or her medical data, the exchange of medical data and the consequences to follow unjust and unauthorized dealing with such data. The Health Insurance Portability and Accountability Act of 1996 (HIPAA, Title II) requires the Department of Health and Human Services (HHS) to adopt national standards for electronic health care transactions and national identifiers for providers, health plans, and employers. The Department of Health and Human Services (HHS) published a final Security Rule in February 2003. The Security Rule specifies a series of administrative, physical, and technical safeguards for covered entities to use to assure the confidentiality, integrity, and availability of electronic protected health information. The Patient Safety and Quality Improvement Act of 2005 (PSQIA) was published on November 21, 2008, and became effective on January 19, 2009. It states the requirement of adoption of national standards for electronic health care transactions and code sets, unique health identifiers, and security. The HIPAA Privacy Rule provides federal protections for personal health information held by covered entities and gives patients an array of rights with respect to that information. At the same time, the Privacy Rule is balanced so that it permits the disclosure of personal health information needed for patient care and other important purposes.

In Europe act the Convention 81/108/EC [3] for the protection of individuals with regard to automatic processing of personal data and Directive 95/46/EC [4] (known as Data Protection Directive) on the protection of the individuals concerning their personal data and the movement of these data. The health and medical data are discussed in Recommendations R(97)5 [5] of the Council of Europe. Directive 2002/58/EC [6] of the European Parliament and of the Council of 12 July

2002 concerns the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications). This Directive is also known as E-Privacy Directive. This is an EU directive on data protection and privacy in the digital age. It presents a continuation of earlier efforts, most directly the Directive 95/46/EC. It deals with the regulation of a number of important issues such as confidentiality of information, treatment of traffic data, spam and cookies. This Directive has been amended by Directive 2006/24/EC [7] on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks. Directive 2009/136/EC [8] of the European parliament and of the Council of 25 November 2009 amends Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services, Directive 2002/58/EC and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws.

On 2 July 2008, the Commission adopted a proposal for a directive of the European Parliament and of the Council on the application of patient's rights in cross-border healthcare. The proposal was sent by the Commission to the EDPS for consultation, in accordance with Article 28(2) of Regulation (EC) No 45/2001. Commission Decision of 22 December 2011 2011/890/EU provides the rules for the establishment, the management and the functioning of the network of national responsible authorities on eHealth. In January, 2012 a proposal for a new legal framework for the protection of personal data in the EU as set out in Communication COM (2012) 9 final [11] was made. This concerns the Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Commission Decision of 18 April 2012 2012/200/EU amends Decision 2004/452/EC laying down a list of bodies whose researchers may access confidential data for scientific purposes (notified under document C(2012) 2291).

In a number of states (Finland [12], New Zealand [13], United Kingdom, Canada [9], Ireland [14]) act different legal issues and regulations.

In Bulgaria the legal issues on privacy and confidentiality of electronically stored patient data are not yet settled. The results from an enquiry carried out among the physicians in Bulgaria in 1999 show that 76.1% of the respondents (medical doctors) are not acquainted with such issues. In a study carried out in 2011 this result is still valid, the majority of physicians 57.6% and 48.7% of nurses are not acquainted with such



issues. The majority of physicians (88.6%) and the majority of nurses (80.9%) consider that in Bulgaria the legal issues concerning privacy and confidentiality of medical data are inadequate. In 1999, 91.8% of the physicians state that special measures are needed concerning the protection of electronically preserved medical data. The results of our study in 2011 show that this result hasn't changed, 85.6% of the physicians and 78.7% of the nurses consider that special measures are needed for the protection of electronically preserved medical data. The measures proposed by the respondents can be combined into several categories: passwords, coding, access codes, digital signature, special legislation, European legislation. These categories are of different nature. Some of them envisage legal issues, which in Bulgaria are not yet resolved. On the other hand, there are security measures concerning information technologies used in these area – digital signatures, access codes etc.

**The security measures** originate along with the confidentiality measures but substantially differ from them. They concern the health system and are related to the responsibilities of the health organization and its functional requirements. The security of health and medical data involves such issues as patient and user identification, access control, availability and reliability, data integrity, responsibility, audit, control, physical security of the equipment etc. The real problem is that most of them go in parallel and depend on the confidentiality measures. In order to introduce proper security measures, all these issues should be initially legally solved.

Besides the question of user identification which solution depends mainly on technologies, the main question that must be solved with utmost attention and precision is the assignment of access rights to health and medical patient data. Access rights should be assigned on the basis of precise analysis of the needs and requirements of the users (user aspect) for access to the different sets of medical data (data aspect) in relation to the healthcare institution (institutional aspect). This requires a proper definition of the levels of medical data for access by the different healthcare providers and other personnel in and out the healthcare institution. We can propose an example of criteria for the access of different healthcare providers to the different types of data in the electronic medical record. The data contained in the electronic medical record may be divided logically to five main categories:

- Unidentified data suitable for different clinical and epidemiological surveys. Although without any personal information, the use of this data should be in respect to the principles of informed patient consent.

- Partially identified administrative/financial data (diagnoses, procedures, treatment, hospital stay) for use by the accountant and administration personnel of the healthcare organization. Access should be granted with respect to the appropriate high level of confidentiality.
- Identified data representing the current health status of the patient (chronic diseases, allergies, risk factors, current treatment, other sensitive information i.e. diagnosis AIDS etc.) that can be accessed by specified healthcare providers in cases of emergency or in emergency settings.
- Identified data concerning current episode of care available to the treating physician and if necessary to other healthcare providers in the same health institution. If not vitally important, access should not be granted to external personnel.
- Identified data concerning the patient status as a whole. Practically this is the entire medical record. Access to this data should be granted only to the patient's personal physician who is responsible for the overall health status of the patient and the patient himself.

The technical solution of the security issues with the help of the modern information and communication technologies is a matter of proper funding and management. Each health organization should assess the technologies for data protection in two aspects: efficiency of the protection of patient's data and the costs of the technologies. Special attention should be paid to such important issues as:

- The restriction or prohibition of physicians' access to information, necessary for medical decision making arising from the patient's right for privacy;
- The purchase of the security technologies in the information system environment;
- Prices for current management, operation and maintenance of the information system;
- Prices for user dissatisfaction generated as a result of improper operating interfaces and procedures;
- Prices for the time spent by the user for fulfilling the security requirements etc.

The contradiction of these aspects is evident. If all security measures are in conformity with the confidentiality measures and fulfill the requirements for protection of patient's rights for privacy, they will be both excessively expensive and restrictive. Therefore the healthcare organizations should aim to implementation of balanced measures concerning the actual threats to information security and the risks against unauthorized acts but on reasonable price. This will acquire appropriate

education of managers, system maintenance personnel and above all users, including the patients themselves.

In general, confidentiality and security measures should be properly balanced. We can point to two different categories of factors that influence both confidentiality and security. On one hand there is the patients' attitude as well as the general opinion, fed by the mass media and other factors bringing up the fear that computerization would violate the patient's rights for privacy. There are still individuals who consider that the paper medical record is much more secure and protected than the electronic medical record. On the other hand, we can point out to at least three issues against confidentiality and security:

- The confidentiality may restrict access to vital medical information and therefore obstruct the delivery of proper healthcare thus causing negative influence on quality of care;
- The strict confidentiality and security measures may prove too hard for the user of medical data;
- Confidentiality and security may require enormous financial and human resources.

Our opinion is that currently it would be better to pay attention mainly to the security policy as hardware measures are still to be improved and security devices are still expensive.

### **Conclusion**

- The emerging world of eHealth involves systems which hold increasingly detailed levels of clinical information that requires special measures for its protection.
- The patient's rights for privacy are in the basis of all confidentiality and security measures.
- The adequate protection of health and medical information depends on the development of an adequate unified patient identifier and the adoption of adequate standards for confidentiality measures.
- The most substantial threat to the privacy of medical data comes from the shared use of this data by different healthcare providers and the inadequate legislation in this field.
- Hardware security measures are an important issue for the protection of patient's rights but special attention should be paid to the security policy.
- Information technologies become even more important for improving the quality and decreasing the costs of healthcare. Therefore the attempts to protect the privacy of patient's medical data should be aimed to finding ways to protect the sensitive medical data, stored and

communicated electronically and not to obstruct the use of information and communication technologies in health organizations.

### References

1. A patients charter for eHealth Information Systems, Prepared by the Patients' and Citizens Working Group of the European Health Telematics Association (EHTEL), Brussels, 2008
2. ASTM, 1994, Membership Information Packet: ASTM Committee E31 on Computerized Systems, Philadelphia, PA
3. Council of Europe Convention No.108 for the Protection of individuals with regard to Automatic Processing of Personal data, Strasbourg, 28 January, 1981
4. Council of Europe Directive on the Protection of Individuals with regard to the Processing of Personal Data, 95/46/EC, L.281/31-50, 23 November 1995. Official Journal of the EC, L-2985, Luxembourg
5. Council of Europe Recommendations R(97)5 on the protection of medical data , signed on 13.02.1997
6. Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)
7. Directive 2006/24/EC of the European Parliament and of the Council of 15 March 2006 on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks
8. Directive 2009/136/EC of the European Parliament and of the Council
9. EU/SO-GIS, document 030/95, March 16, 1995, the case of Canada
10. [http://ec.europa.eu/information\\_society/eeurope/i2010/index\\_en.htm](http://ec.europa.eu/information_society/eeurope/i2010/index_en.htm)
11. Safeguarding Privacy in a Connected World – A European Data Protection Framework for the 21st Century, COM(2012) 9 final.
12. The Finish Act on Patients' Legal Status and Personal Rights, 1993
13. The Privacy Act, New Zealand, 1993
14. The ePrivacy Regulations 2011 (S.I. 336 of 2011)
15. U.S. Senate. S 1360: Medical records Confidentiality Act of 1995, 1995

### Author:

**Iskra Mircheva**, PhD, Assoc. prof.  
Head Dpt Social medicine and healthcare organization,  
Faculty of Public Health, Medical University of Varna, 9002 Varna,  
55 Marin Drinov Str.,  
Phone: 0887618180, email: [mircheva\\_is@abv.bg](mailto:mircheva_is@abv.bg)

## **CROWDSOURCING – A BRIDGE BETWEEN ACADEMIA AND COMMERCE**

Evgeniya Rakitina-Qureshi, Blagoy Klimov, Gergana Zhelyazkova,  
Krasimir Ivanov, Ivo Rakitin

**Abstract:** *Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. In other words, by using the crowdsourcing it is possible to engage a large number of people, from various fields of knowledge on the development of creative solutions for the problems displayed.*

**Keywords:** *Crowdsourcing, online platform, open innovation, survey, academia, commerce, collaboration.*

### **1. Crowdsourcing Vs Open Innovation**

#### **Open Innovation**

Open innovation is a paradigm that companies can benefit from the use of external ideas as well as internal ideas, to innovate. The boundaries between a company and its environment have become more permeable; innovations can easily transfer inward and outward. The use of open innovation platforms represents a benefit for both parts involved because it provides innovative solutions for the problems posed and gives the opportunity to the solvers to get in touch with the real world problems and challenges.

#### **Crowdsourcing Vs Open Innovation**

Jack Hughes, on an article called The Evolution of Crowdsourcing: Open Innovation, published online on Top Coder (<http://www.topcoder.com/blog/the-evolution-of-crowdsourcing-open-innovation/>) defines Crowdsourcing as “the practice of engaging someone to do a task for you through access to external people (the “crowd”) while Open Innovation is the practice of finding a way to do something new through access to people you normally wouldn’t be engaged with”. According with the same author Open innovation is a larger concept and a “superset of Crowdsourcing – when the process and platform include finding a way to create something new”.

#### **Collective Intelligence and Crowdsourcing to strengthen Research Value Chain**

An important added value related to Crowdsourcing has come from the article of Thierry Buecheleretal, about the Crowdsourcing, Open Innovation and Collective Intelligence applied in the Scientific Method. The article has stressed the importance to enlarge the process of research innovation to think up new inventions. Crowdsourcing can be defined in fact as occurrences of Collective Intelligence which involves

“groups of individuals collaborating to create synergy, something greater than the individual part (Castelluccio 2006). Crowdsourcing applied to fundamental research and private sector is described as a more effective research collaboration “that radically enlarges the pool of (potential) scientific collaborators”.

UC-CROWD/Universities-Companies Crowdsourcing is aimed at strengthening linkages between Companies and Higher Education Institutions (HEI) by mean of the reinforcement of an alternative way of dialogue focused on company's needs. Core aim of the project is to contribute to “new learning and teaching methods” by creating a Crowdsourcing platform to match company needs and academic innovative solutions. The platform will highlight all the “challenges” (problems to solve) provided by companies specialized, in this pilot version, in the energy sector. University professors and students will further reply to the challenges suggesting their “solutions”. The international “crowd” is the real protagonist of this innovative project venture. An European and International base of enterprises' and universities' stakeholders will contribute to the empowerment of the UC-CROWD Platform to create a unique network oriented towards energy efficiency and sustainable growth.

In order to strengthen the knowledge triangle between education, research and business, the survey has been designed to investigate the actual expectations and needs related to the Crowdsourcing platform stressed by the main project actors: Companies, Professors and Students. In particular through the survey it will be possible to tailor the platform matching the real needs and expectations. As an example, responses will highlight services should the platform offer to be attractive, what benefits companies could have from being closer to the HEI's and vice versa.

## **2. Professors' Survey Background**

UC CROWD survey has been designed and developed to identify and process data about Crowdsourcing competence, background motivations, obstacles, incentives or relevant functionalities witnessed by entrepreneurial target group.

The report furthermore highlights the emerging strengths or threats that could positively or negatively influence the project's success.

As previously mentioned, one of the most important project stakeholders' group is represented by the professors which will supervise the quality of the provided solutions to the real life projects in the Crowdsourcing Platform.

A suitable involvement in the project will influence the next success of the project. Through responses professors have pointed out actual

competencies in the field, perceptions on how the platform should be to cope with their needs and expectations.

The Survey has been designed and developed by partners. In order to facilitate the management and the administration of the questionnaire a unique survey has been designed addressed to all 3 target groups: companies, university professors and university students.

Although there was a unique survey, specific questions were directed to investigate the specific needs of professors.

Survey has been uploaded into online survey software in order to facilitate the collection and the elaboration of data.

As concerns the statistic research strategy, the questions with rank responses (1-7) have been analysed providing:

- the visual representation of **row data** related to each option which allows you to understand the amount of responses for each option;
- the visual representation of **means** related to each option which allows to point out immediately the most important options chosen by participants
- Finally, the visual representation of the **confidence intervals** related to each option, a type of interval used to indicate the reliability of an estimate.

Beside that, a grid for each matrix question provides all the aforementioned values including the **standard deviations** which show how much variation or dispersion from the average exists.

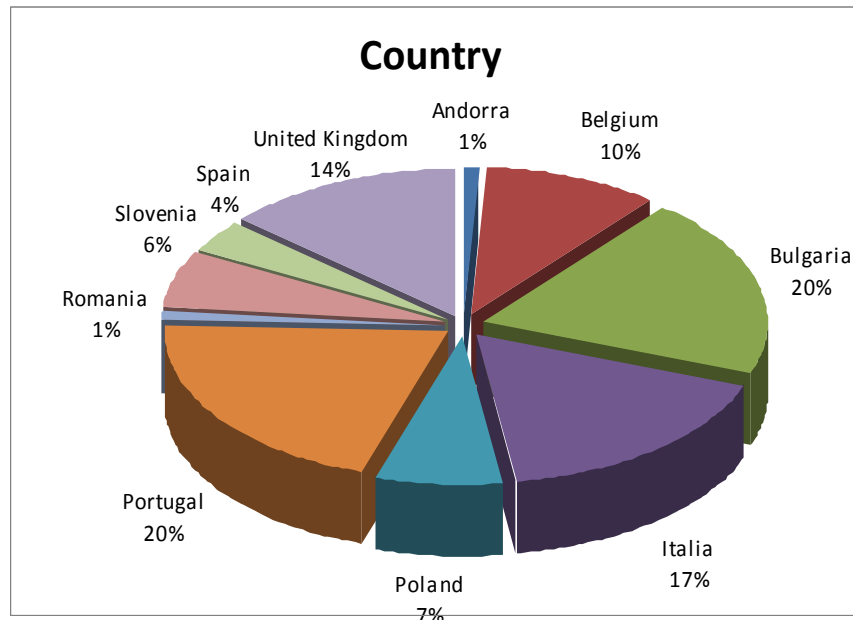
### 3. Analysis

The survey reached a cluster of 111 professors coming from European and Third Countries:

#### 3.1. European Countries Questionnaires

**Table 1**

Andorra	1
Belgium	11
Bulgaria	22
Italia	19
Poland	8
Portugal	23
Romania	1
Slovenia	7
Spain	4
United Kingdom	15



**Figure 1**

### **3.2. Section Motivations**

**Purpose:** Matrix Question was addressed to understand the factors that could mostly motivate the **professors** to actively participate to a Crowdsourcing platform.

**Analysis:** as concern motivation factors the most relevant motivation engines are represented by

1. Creating contacts with stakeholders e.g. companies / academics / communities / students / research centers.
2. To transfer scientific knowledge into practice, by developing research projects based on the companies' problems
3. The satisfaction of solving of a problem or receiving a solution for your problem
4. New ideas for research opportunities

The responses provided witness that the **professors** representatives have very clear priorities. The first is to create new contacts and maintain relationships with relevant stakeholders. The second is to transfer scientific knowledge into practice, by developing research projects based on the companies' problems. The third important motivation for professors is the satisfaction of solving of a problem, using their knowledge and experience.

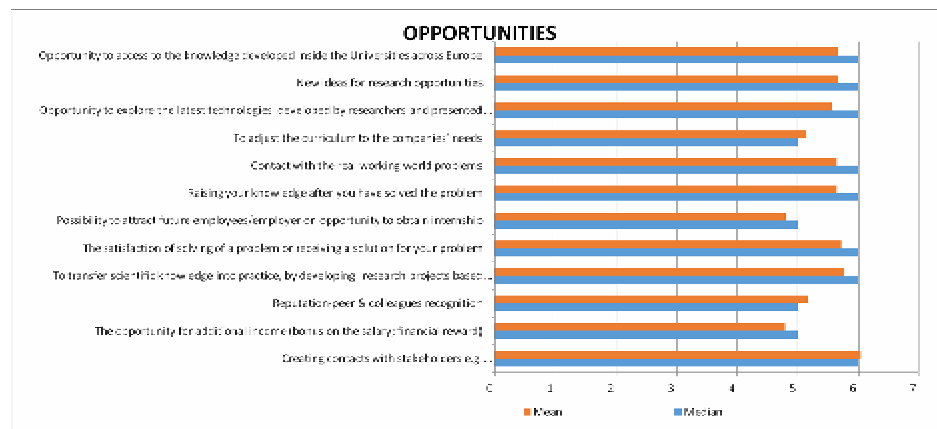


According to these important lessons acquired it is necessary for the UCCROWD project that its CS Platform will be highly diffused and benefit from many participants. Academic and business Professionals need to join the platform in order to enrich the level of scientific exchange.

The second important lesson relevant for the project is that the level of qualification of participants needs to be high to be able to provide successful solutions of real life problems and to transfer scientific knowledge into practice.

In general, all the participants have chosen high rates witnessing that, on average all the responses have been considered relevant. Very few have chosen from 1-3 of the rank.

Confidence intervals related to each option indicate the reliability of the estimate.



**Figure 2**

### 3.3 Section Obstacles

**Purpose:** Matrix Question was addressed to point out the factors that could damage and negatively influence the development of a crowdsourcing platform.

**Analysis:** as concern obstacles the most relevant threats were represented by:

1. Intellectual Property issues especially on the ownership of ideas (copyright, patent etc.)
2. Companies have the perspective that there is a gap between what is taught in universities and what is useful for companies.

3. Difficulties of companies to internalize the knowledge from outside i.e. it may be difficult for companies to implement external help/ideas/assistance/etc.

4. Lack of participation from the companies

The responses provided witness that the **professors** representatives have a main concern related to the Intellectual Property issues especially on the ownership of ideas. According to this information UCCROWD platform needs to provide solutions to protect copyright owners in order to encourage professors to share their intellectual property.

Beside that, the professors' concerns are that companies have the perspective that there is a gap between what is taught in universities and what is useful for companies. This could be based on professors' previous experiences with the business.

Academic is concerned that it may be difficult for companies to implement external help/ideas/assistance which could create possible threats for project success. Professors' workshops could be a profitable ground to establish linkages and contrast the aforementioned perspective. Face to face communication during such events could change the past experiences.

Professors are also concerned that there could be lack of participation from the companies, which would be a significant obstacle for the success of the platform.

Finally it is interesting underline that the less relevant obstacles were that the platform was in English and the exchange ground was through an online platform. The basic knowledge related to English language and familiarity with platform represents positive aspects for the widespread use of such environment.

Confidence intervals related to each option indicate the reliability of the estimate.

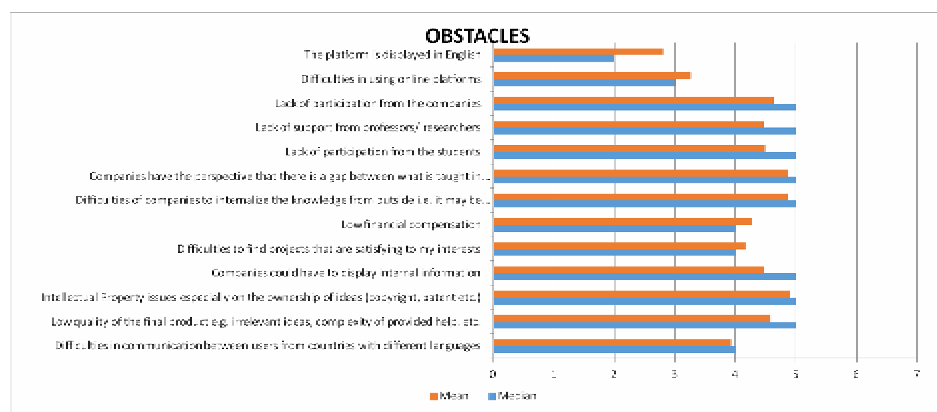


Figure 3

### 3.4. Section Incentives

**Purpose:** Matrix Question was addressed to point out the incentives (in particular academic) that could be used to encourage professors to participate with the crowdsourcing platform.

**Analysis:** as concerns incentives the most relevant opportunities are represented by

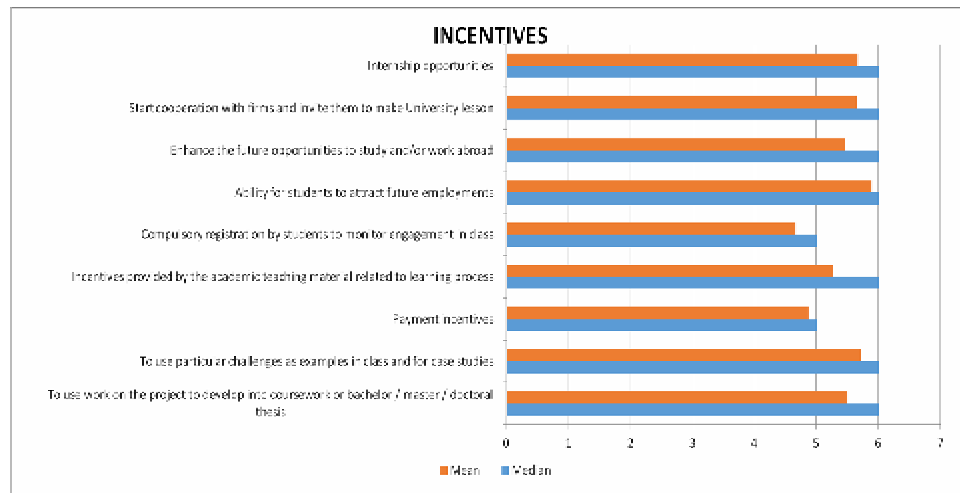
1. Ability for students to attract future employments
2. To use particular challenges as examples in class and for case studies
3. Internship opportunities
4. Start cooperation with firms and invite them to make University lesson

Responses witness that professors believe that the biggest incentive is the ability for the students to attract future employments.

Professors believe that the projects provided through the crowdsourcing platform could be used as examples in class and for case studies, which eventually will increase the practical side of the lessons.

Professors also find important the Internship opportunities and the opportunity to start cooperation with firms and invite them to make University lesson.

Confidence intervals related to each option indicate the reliability of the estimate.



**Figure 4**

### **3.5.1. Basic information of users**

Purpose: To understand the proposed functions that were suggested and preferred by the respondents so that it can incorporate it into the platform.

Analysis:

Most of the information mentioned in the questionnaire related to the Student, Professor and University profiles has been assessed as relevant; they only recommended not including were i.e. the title, the address. .

The same, in general, was for the company profile where they only recommend not i.e. including the title, the amount of transactions.

It is suggested including in the registration some “compulsory” information, the other as optional.

The other suggestion is to exploit the university page of the department/professor for the academic staff.

### **3.5.2. Functionality**

Purpose: To understand which are the most important functions suggested and preferred by the respondents so that it can incorporate it into the platform.

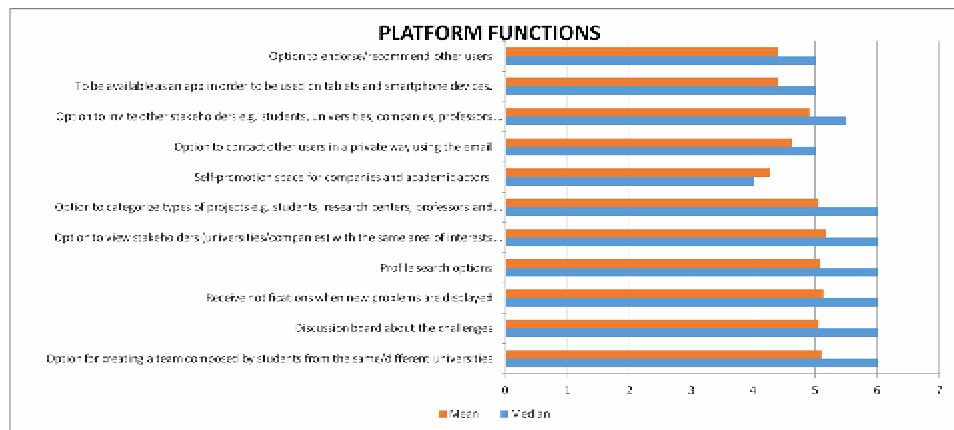
Analysis: as concerns the platform functions, in general, all the functions have been considered relevant and necessary. In particular the first places underlined as the most important have been the following:

1. Option to view stakeholders (universities/companies) with the same area of interests Projects that are satisfying my scientific fields
2. Receive notifications when new problems are displayed
3. Option for creating a team composed by students from the same/different universities
4. Profile search options

According to the professors opinion the most important Functionality of the platform is the Option to view stakeholders with the same area of interests. Projects that are satisfying their scientific fields, which express their interest to form teams to work on different challenges.

Also very important Functionality for professors is to receive notifications when new problems are displayed. This means they want to be informed on any new challenge, which could provide them an opportunity to use their knowledge to solve the real-life problems.

Professors feel that the Option for creating a team composed by students from the same/different universities has significant value and also a Profile search options included would improve their benefits from the platform.



**Figure 5**

#### **4. Conclusions**

According to the professors' responses, as concerns INCENTIVATION, professors would be motivated knowing that the platform could provide the ability for the students to attract future employments. Also it is very important for the professors to be able to use particular challenges as examples in class and for case studies.

According to these important lessons acquired it is necessary for the UCCROWD project that its CS Platform to be highly diffused and benefit from many participants. It is very important for the Academic platform to provide benefits for the students both in finding future employment and in solving real time problems in the classes.

The next important feature of the platform from professors' point of view should be providing the internship opportunities and development of cooperation with firms and involving them in university lessons.

According to the professors' responses, as concern OBSTACLES, professors expect difficulties with protecting their intellectual property and especially on the ownership of the ideas. According to this information UCCROWD platform needs to protect the copyright owners and to provide a solution for Intellectual Property issues.

Moreover, professors believe that the Companies have the perspective that there is a gap between what is taught in the universities and what is useful for the companies. Professors are worried that this could reduce the success of the platform.

Professors think, based on their previous experience that there are some Difficulties of the companies to internalize the knowledge from outside i.e. it may be difficult for the companies to implement external help/ideas/assistance/etc. To ensure the success of the platform a

dialogue should be guaranteed between academic and companies through workshops and also through special features of the platform such as forums.

As concerns FUNCTIONS, professors ask the possibility to view stakeholders (universities/companies) with the same area of interests Projects that are satisfying their scientific fields. Very important feature of the platform is the possibility to Receive notifications when new problems are displayed and the Option for creating a team composed by students from the same or different universities.

1. **Evgeniya Rakitina-Qureshi**, Associate Professor, PhD, rakitina@vfu.bg, +395 887 498 358
2. **Blagoy Klimov**, PhD, blagoy.klimov@gmail.com, +359 884 610 398
3. **Gergana Zhelyazkova**, Assist. Professor, gzhelyazkova@yahoo.com, +359 882 415 729
4. **Krasimir Ivanov**, Assist. Professor Eng., krassswr@gmail.com, +359 887 380 054
5. **Ivo Rakitin**, Eng. Inf., ivo1st@gmail.com, +359 887 374 740

# ARCHITECTING SYNERGY BETWEEN TECHNOLOGY-BASED INNOVATIONS FOR E-GOVERNMENTS AND E-GOVERNANCE TO ENABLE BUSINESS INNOVATIONS

Alexander Samarin

**Abstract:** *Experience has shown that the stand-alone deployment of individual technology-based innovations usually brings relative improvements but it may also bring unexpected new challenges. The latter are often caused through the sub-optimal use of a particular technology that promises, or has been forced, to solve “all” existing problems within an organisation. Experience has also shown that synergy between various technologies does not happen spontaneously; it must be architected for guiding the use together of new technologies.*

**Keywords:** *enterprise architecture, business architecture, BPM, SOA, ECM, IT governance, information security, e-government, e-governance*

## 1. Introduction

Looking at e-governments from a global perspective, we can conclude that all e-governments deliver the same services, albeit in a different manner. The natural differences (based on the core businesses) are found in the languages, governance structures, laws, distribution of power, social support practices, etc. The implementation differences occur in the use of different tools, different skills, different architectures, different service providers, etc.

Certainly there is a cost behind such differences. At present, this is the cost of implementing about 200 e-governments in about 200 different ways, even if there are many commonalities among implementations of e-governments. Let us estimate that the natural differences are only 20 % of the total cost of ownership of the e-government (this estimation is based on the author's experience with various e-government projects). This 20 % must be implemented separately for each e-government; the remaining 80 % can be done once for all e-governments, with very high quality and in close cooperation.

Because of the potential considerable gain in cost, time and quality, it is worth asking the question of how to combine diversity (i.e. endorsing or appreciating the natural differences) and uniformity (standard/common solutions wherever these bring value).

This article is aimed at outlining how to architect synergy between various modern methodologies and technologies to implement e-government and e-governance combining both diversity and uniformity. Enterprise Architecture (EA) is the right “tool” to address this question because EA is a systemic integrator of people, processes and technologies in 4 dimensions (scope, architecture, time and sectors).

This article proposes three conceptual pillars:

- an e-government reference model for achieving a common understanding,
- an e-government platform-based architecture for agile development of potentially shared solutions and
- explicit representation of a governmental entity as a system of processes for the reduction of undesired complexity.

## **2. E-government reference model**

The e-government reference model is fully presented in [1]; the key views of this reference model are discussed below.

### **2.1. Partner – government interaction view**

Partner – government interaction is considered as a long-running process with participation of a partner (citizens, local businesses and other organisations) and several governmental entities (ministers, agencies, local communes, etc.). To protect the partner from the internals of the government and to unify his/her interactions with the government, the e-government acts as a shell which coordinates the flow of data (and documents) between the partner and the governmental entities.

To streamline the work of partners with the government, e-government provides a social collaborative extranet for all partners. This extranet:

- helps partner to manage all electronic documents (which are exchanged between the partner and the government) in a secure manner;
- helps partner to execute his/her tasks, and
- allows a partner to interact with other partners.

This extranet is an interface layer between the partners and e-government (thus with the whole government as well) as shown in figure 1.



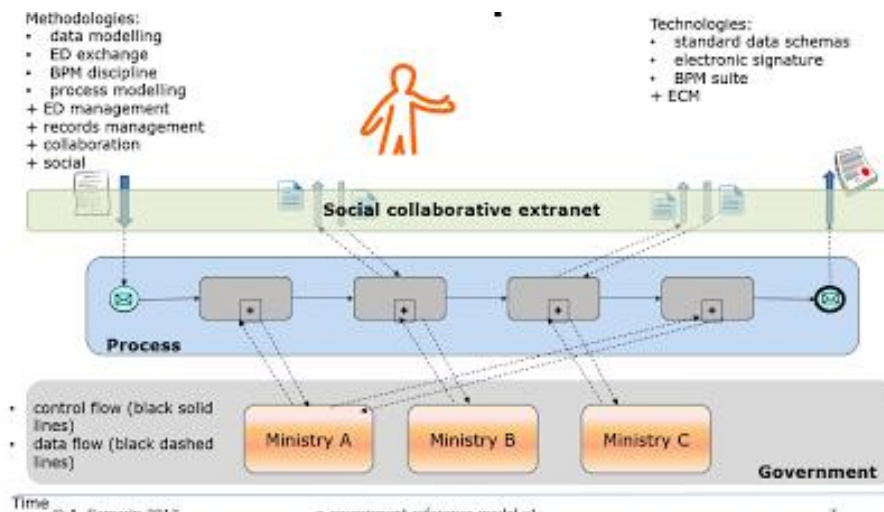


Fig. 1. Position of the social collaborative extranet

## 2.2. Partner view

For the partner, the social collaborative extranet may have the following visual design (see figure 2).

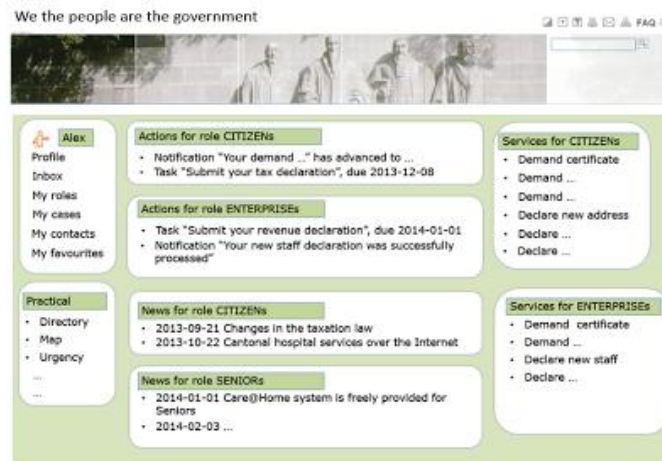


Fig. 2. Example of the partner's view

This extranet considers that:

- Each partner has several roles, e.g. YOURSELF (person and his/her legal representatives), CITIZEN (person officially leaving in this

country), ENTERPRISE (manager of a business), SENIOR (persons with age over 60 years), etc.

- Person may select which roles he/she is carrying out at a particular moment in time.

- Each communication between a partner and the government is a case with associated documents, data, audit trails, records, service level agreements and key performance indicators.

- A case may be completed or on-going.

### 2.3. The governmental entities integration view

E-government, which acts also as an inter-ministerial coordination tool, ultimately resolved the integration problems within the government. Instead of that ministries are connected to each other in ad-hoc way, the e-government offers an integration process which delivers data and documents between all ministries via a centralised service (backbone) for the inter-ministries secure electronic exchange (like sending / receiving registered letters).

### 2.4. Application architecture view

In general, the e-government application architecture is as in figure 3. There are three main technologies:

- Enterprise Content Management (ECM) for the social collaborative extranet;

- Business Process Management (BPM) for coordination and integration backbone, and

- Service Oriented Architecture (SOA) for coordination and integration backbone.

- 

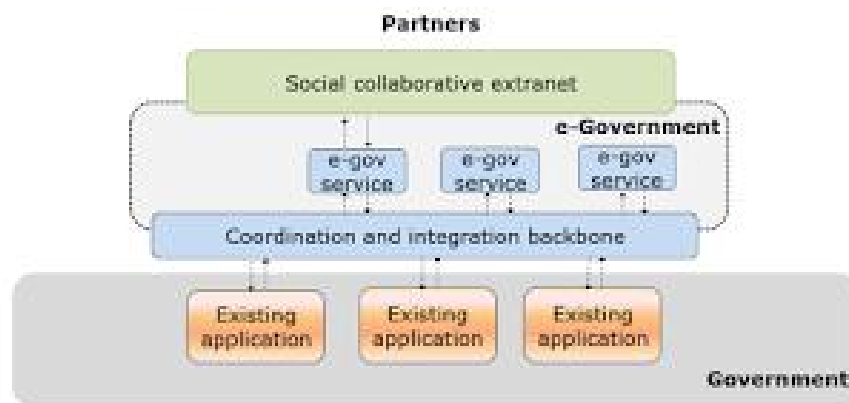


Fig. 3. Application architecture overview

Actually, this architecture is only a stage in sort of ladder for step-by-step evolution as shown in figure 4.

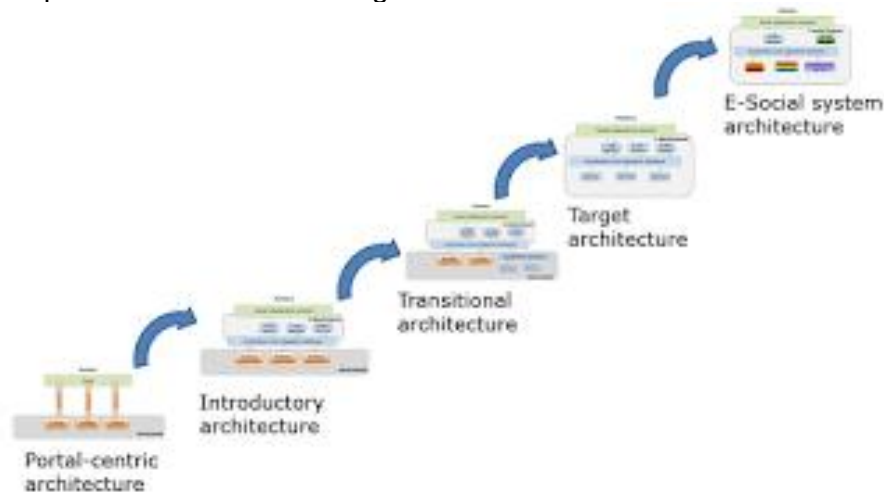


Fig. 4. Application architecture – all stages as a ladder

## 2.5. Use of SOA and BPM

The application architecture should be based on BPM and SOA. From the business point of view on BPM/SOA, an executable process coordinates some services as shown in figure 5.

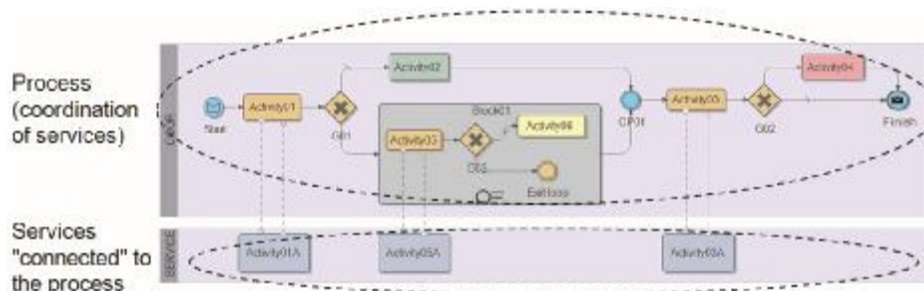


Fig. 5. BPM and SOA are working together to have executable processes

But from the IT point of view, there may be many different services around each process, because a process requires various artefacts: events, roles, rules, documents, data, audit trails, and performance indicators. To help structure different services and other artefacts around processes, a multi-layer implementation model (see figure 6) is proposed

for BPM/SOA solutions. In this model, each layer is a level of abstraction of the business and addresses some particular concerns.

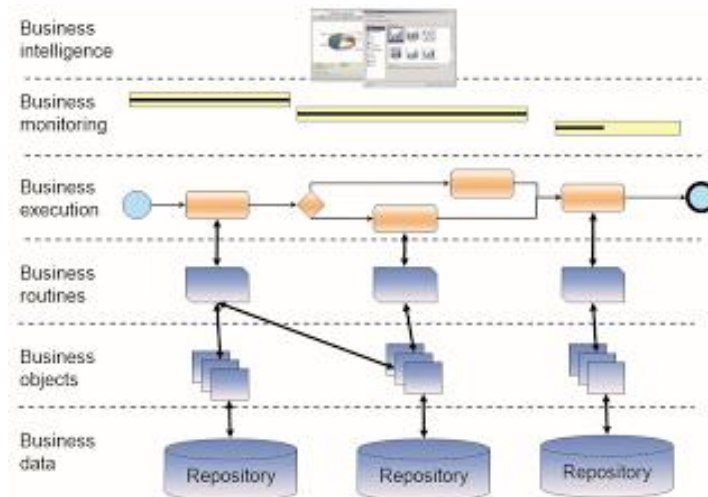


Fig. 6. Multi-layer implementation model for BPM/SOA solutions

### 3. Platform-based architecture

The platform-based architecture is fully presented in [2]; some key architectural principles of it are discussed below.

#### 3.1. Platform description

The platform offers various ready-to-use and coherent technical, functional, and business capabilities optimised for typical public sector activities: education, order enforcement, civil protections, healthcare, finance, people benefits and rights, land, natural resources, legal system, governance, local business enabling, etc.

To cover variations between different countries, the platform will enable a high level of adaptability to

- policies and priorities of the government,
- structure of ministries and departments,
- legislation,
- operational business processes,
- available budget, and
- level of computerisation.

As those public sector activities share many common capabilities, the latter can be implemented as a platform which serves all activities

(see figure 7). This architectural pattern combines the advantages of the platform and opportunities for quick delivery:

- standardise and simplify core functionality as a coherent foundation and
- speed up the implementation of new innovative solutions on top of the foundation.
- 

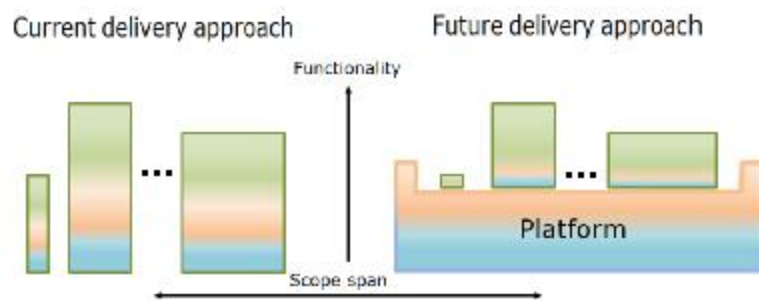


Fig. 7. The foundation architectural pattern

The platform is a coherent set of governance rules, architectural principals, best practices, integration patterns, shared ICT tools and solutions. Key characteristics of the platform are:

- incremental creation and gradual deployment;
- high maturity level of corporate basics, e.g. with records management, risk management, electronic management of documents, and management by processes;
- high maturity of IT operations, management and governance;
- everything is transparent and traceable (thus trustful):
  - design of e-government as a system
  - execution e-government and e-governance services
  - management of data & document repositories
  - some data are open
  - evolution of e-government and e-governance
- risk-awareness (including proactive risk evaluation) is built-in by design;
- paperless exchanges between government, citizens, business, and other partners from the civil society;
- on demand involvement of remote experts (external oversight);
- government operational excellence;
- made and run by national ICT companies.

Being matured and gradually enriched, the platform will enable the agile implementation of e-government and e-governance services in

regional countries. Also, the platform will guarantee the seamless integration and interoperability of those services on national and regional levels.

### 3.2. Platform implementation

The implementation is fundamentally incremental. Platform's architecture is built incrementally in the top-down manner, tools are incorporated incrementally by demand, regional countries are addressed incrementally in accordance with their priorities, business domains are covered as demanded, and different components of the platform are implemented as typical development projects. At the same time, architecture, tools, and other components are re-usable among participating countries.

An ideal implementation scenario would be:

1. Quick development of the architectural foundation for the platform.
2. Identification of current development projects.
3. Guidance for those projects.
4. Coordination and evolution of the platform's architecture.
5. Repeating steps 2-4.

The use of the benefits of the platform depends on the readiness of each participating country to advance with its suitable pace. A “ladder” or “maturity model” (Figure 8) is a metaphor for suggesting how a set of countries with different abilities might achieve common goals and plan their progress. The “ladder” has a few levels of capability from “not able” to “fully capable”. The possibility for participating countries to advance to the same direction at their own pace is guaranteed by the architecture of the platform.

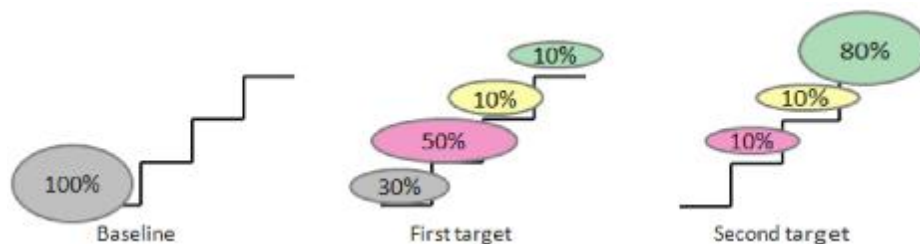


Fig. 8. Maturity ladder

The platform is built from commercial and open-source software (Figure 9). Free and open-source software is used by preference. The global partnership with big IT vendors is to be established.

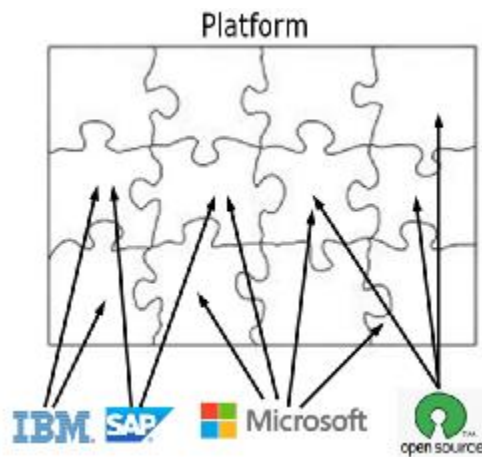


Fig. 9. Platform implementation from different sources

A strong and robust eco-system for start-ups, innovation clusters and centres of excellence should be created around the platform.

#### **4. Governmental entity as a system of processes**

A view of an enterprise (thus any governmental entity as well) as a system processes is fully presented in [3]; the key points of this are discussed below.

##### **4.1. Processes as a business “tool” to reduce the complexity of an enterprise**

Enterprise functioning can be considered as business activity flows spanning the applications, employees, customers and partners within and beyond the boundaries of the enterprise. (Business activity is a unit of work). Individual business activities are interrelated statically and dynamically.

Although any enterprise has its naturally complexity, the lack of explicit and codified knowledge about all relationships adds a lot of undesired complexity. The latter is the root cause why it is difficult to manage and improve enterprises.

Historically, business process is the basic business concept to bring some order into the structure and the behavior of interrelated business activities. A business process is explicitly-defined coordination for guiding the enactment of business activities.

In its simplest form, the explicitly-defined coordination is an initially agreed plan to follow a defined sequence of activities; the plan may

include some variants and the plan allows some changes during its execution. A detailed, formalised and unchangeable plan (e.g. in a form of process template expressed in BPMN) is one of the several coordination techniques (see [4]). The explicitly-defined coordination reveals the structure and the behavior and thus reduces the undesired complexity.

In general, process templates must not be considered as something which is “fixed forever”. As shown in [5] relationships between process template and its process instances can be various.

Because the coordination can be strong (e.g. as in the army) or weak (e.g. as in an amateurs football team), it is possible to discover various constructs (in addition to business processes) within the whole structure and behaviour of business activities within an enterprise. Usually, the coordination between business activities which belong to a particular construct is stronger than the coordination between similar constructs. The number of relationships between business activities which belong to a particular construct is higher than the number of relationships between business activities which belong to different (but similar) constructs.

Thus, the knowledge about constructs reduces the number of potential relationships between business activities and allows concentrating on constructs themselves and relationships between constructs.

#### **4.2. Coordination within processes**

It was observed that although all business processes are unique, they are actually composed from similar process fragments or process patterns. For example, let us consider a typical “repair” process in housing management which comprises claim making, evaluation, selection of a service provider, repair, control, invoicing and insurance to pay the activities. This process several is actually composed from a few process patterns as shown in figure 10 (an original process model is covered by patterns).



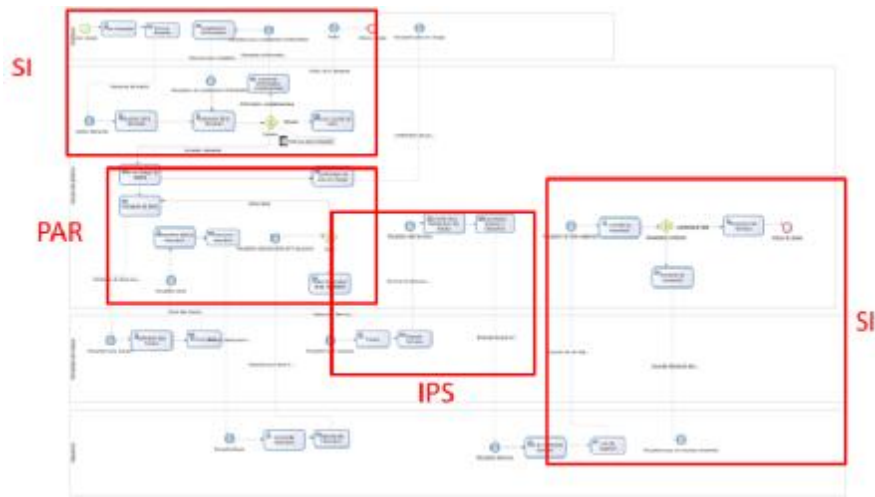


Fig. 10. Process with process patterns

Patterns used in this illustration are:

- Pattern SI – “Submission Interface”
- Pattern PAR – “Propose, Act and React”
- Pattern IPS – “Initial Process Skeleton”

Although any pattern is a highly-cohesive construct, patterns can be adapted to particular needs. For example, the Delegation of Authority Matrix (DAM) pattern is implemented differently in the procurement and in the project management.

A collection of practical process patterns for quick and systematic implementation of unique business processes is available in [6].

#### 4.3. Coordination between processes

The coordination between processes is usually event-based to implement complex interactions between processes as show in figure 11.

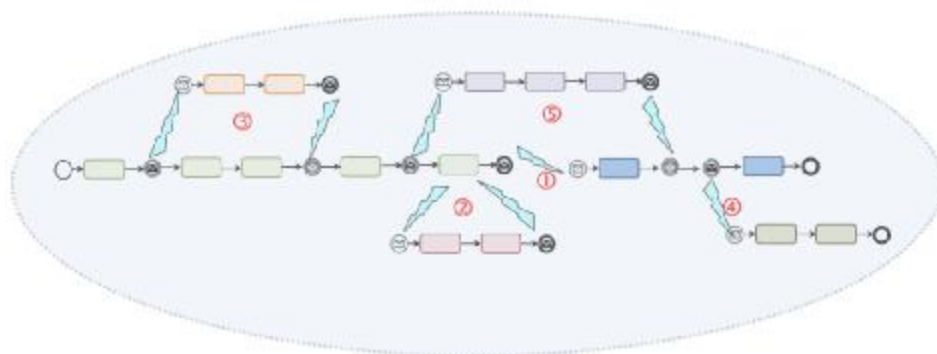


Fig. 11. Examples of coordination between processes

- the finish of one process is the start of another process – marked with circled “1”;
  - starting another process and just waiting for its finish (aka synchronous invocation) – marked with circled “2”;
  - starting another process and waiting for its finish sometimes later (aka asynchronous invocation) – marked with circled “3”;
  - starting another process and not waiting for its finish at all (aka “fire and forget”) – marked with circled “4”;
  - starting another process from one process and waiting for its finish in third process – marked with circled “5”, and
  - co-processes (e.g. a pair of rock climbers tied together by a rope agree to climb one after the other, and not simultaneously as shown in [7]).
- Additional coordination techniques are also necessary because:
- some event are generated outside of processes (typical case is a database trigger which is the data-base coordination technique), and
  - events are filtered between processes as shown in [8].

#### **4.4. Cluster of processes**

Processes, which are strongly related to each other, form Cluster Of Processes (CLOP). These processes are usually functional processes which implement a particular business function, e.g. Field Services. There is also a set of related processes which helps to execute a particular CLOP; the former are monitoring, operating and governance processes for the latter. These related processes are also called leading processes.

- Monitoring processes are responsible for analysing the running functional processes – some sort of operational intelligence.
- Operating processes are used for implementing operational (via available parameters or controls) changes.
- Governance processes are used for detecting, designing and implementation of structural changes, e.g. “customer satisfaction and loyalty development”, “quality”, “strategy and planning” etc.

All together, they constitute two control loops – operational and strategic as shown in figure 12. Sometimes, extra health-check-points may be introduced for better monitoring of complex clusters.

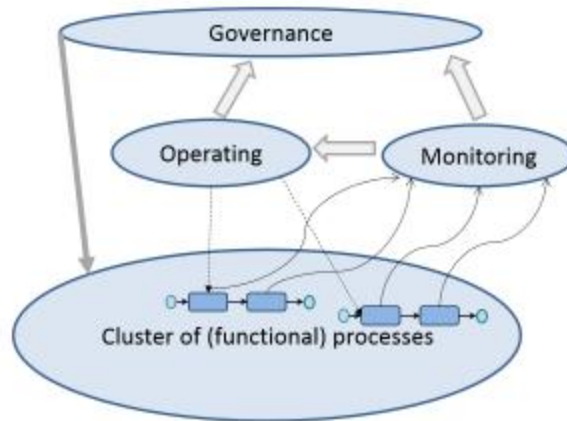


Fig. 12. Functional and other processes

#### 4.5. Coordination between CLOPs

Let us look at situations when some activities from a CLOP initiate (maybe even indirectly) activities in other CLOPs. The explicit coordination via event-based techniques is used in the form of “fire and forget”. For example, a service technician found that a particular piece of equipment must be replaced and he/she has to initiate an order to carry out this replacement.

The implicit coordination between CLOPs may happen as well. There are many situations in which one process is changing a business object and some non-trivial actions may be taken because of those changes. Also, a chain of changes may happen: a trivial change in a business object necessitates a change in another business object which necessitates a non-trivial action. And the business objects maybe of different types.

For example, a field service technician updated (in an activity in the process “close services”) his/her work order (business object 1) with a report which contains some lessons learnt. Those lessons learnt, after some validation, may be incorporated into the product maintenance guide (business object 2) or to the customer’s profile (business object 3). An analysis of customer’s profiles may reveal a need for some preventive maintenance (business object 4) or an opportunity for a custom project (business object 5).

Figure 13 shows the separation between the update logic (change the state as the result of an external stimulus) and life-cycle logic (performing some activities that are related to a particular state).

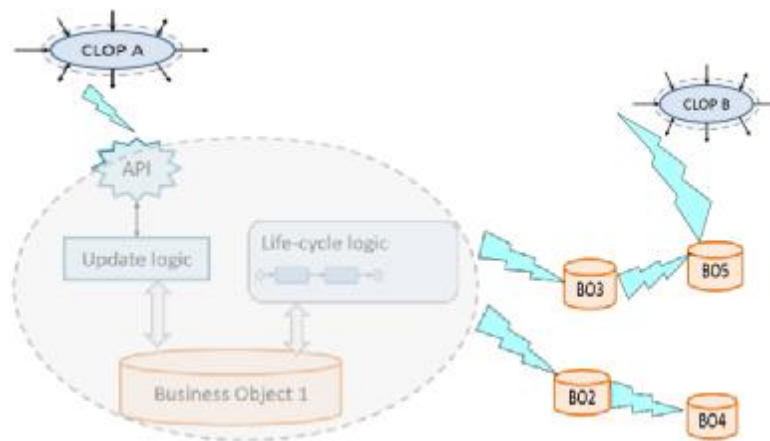


Fig. 13. Implicit relationships between CLOPs

The use of explicit and executable processes, other constructs and relationships between them reduces the undesired complexity of an enterprise, improve understanding of its structure and behaviour, thus facilitates the management and evolution of the enterprise.

## 5. Applying the e-government reference model

Imagine, that a country wants to implement its e-government and e-governance correctly. One of the critical measures of implementation "correctness" is minimal (ideally, total avoided) duplications in its e-government and e-governance implementations while following the goals and priorities of the country.

Typical country has many governmental entities which are spread in the following levels of government:

- National (or federal);
- Federal ministries and agencies;
- Regional (or cantonal or provincial) authorities with their ministries and agencies;
- Districts (and sometimes sub-districts) authorities, and
- Municipal authorities.

As estimated in [1], technical and data architectures are practically common for each level. Application architecture is shared at about 80-90 %. Business architecture is shared at about 70 %.

With this amount of similarities, there is only one "classic" enterprise architecture project is necessary. Such a project must be

carried out on the national level. Other projects are just solution-architecture projects to develop common components.

The e-government reference architecture is the way to bring the conceptual integrity. The latter is mandatory to reach some coherence in decisions about:

- Optimisations for users (external and internal).
- Compliance with laws.
- Avoiding duplication of work among ministers, provinces, municipalities.
- Simplification of evolution of applications.
- Realisation of the national strategy.
- Natural evolution of e-government.
- Standardisation of business processes fragments.
- Unification of software.
- Introduction of new tools and technologies.

The consistent use of BPM, SOA and ECM allows addressing various e-government concerns and removing many duplications in a systemic, technology-independent manner within the whole country and, potentially, among several countries.

## **6. Conclusion**

The benefits of this architecture-led implementation and evolution of e-government are multiple:

- Conceptual integrity of the e-government as a system which allows better decision-making on investments, technology and evolution.
- Progressive fusion of disjointed and duplicated solutions and systems to a coherent platform for business execution, which will act as a catalyst for business innovations.
- Information security is explicit (via processes) thus enabling better change impact evaluation.
- Possibility for worldwide and regional cooperation in the development and re-use of e-government components and solutions.
- Built-in ability to communicate between various governmental entities within and beyond the country.

## **References**

1. "E-government reference model" see <http://www.samarin.biz/pubs/e-gov-reference-model-v1.pdf>
2. "Regional platform for e-governments and e-governance" see <http://www.samarin.biz/pubs/e-gov-platform-architecture-v2.pdf>

3. <http://improving-bpm-systems.blogspot.com/2014/03/enterprise-as-system-of-processes.html><http://improving-bpm-systems.blogspot.com/2014/03/enterprise-as-system-of-processes.html>
4. <http://improving-bpm-systems.blogspot./2014/03/coordination-techniques-in-bpm.html><http://improving-bpm-systems.blogspot./2014/03/coordination-techniques-in-bpm.html>
5. <http://improving-bpm-systems.blogspot.com/2010/12/illustrations-for-bpm-acm-case.html><http://improving-bpm-systems.blogspot.com/2010/12/illustrations-for-bpm-acm-case.html>
6. <http://improving-bpm-systems.blogspot.com/search/label/practical%20process%20patterns><http://improving-bpm-systems.blogspot.com/search/label/practical process patterns>
7. <http://www.slideshare.net/samarin/process-practical-patterns-si><http://www.slideshare.net/samarin/process-practical-patterns-si>
8. <http://improving-bpm-systems.blogspot.com/2011/01/explicit-event-processing-agents-in.html><http://improving-bpm-systems.blogspot.com/2011/01/explicit-event-processing-agents-in.html>  
<http://improving-bpm-systems.blogspot.com/2011/01/explicit-event-processing-agents-in.html>

**Author:**

**Alexander Samarin**, PhD, independent consultant, Switzerland,  
alexander.samarine@gmail.com, +41 76 573 40 61

# **INFORMATICS IN ENGINEERING AND TECHNOLOGIES**

## **UNIQUENESS OF THE IMAGE LIGHTER IN SIZE FOR WEB SPACE**

Mariya Deshkova

**Abstract:** *The report presents vital to the creation of a unique pictorial several programs to work with graphics, as well as opportunities to reduce file size with a view to their rapid publication WEB space*

**Keywords:** *uniqueness, pictures, photos, ways to unify, image resizing, file size, Web space*

To make a website interesting and often visited, it must have unique content and unique images. We cannot imagine the sites without media - images, audio, video. They would be completely uninteresting. Here I want to tell you in detail about the pictures, and in particular how to achieve uniqueness of the picture and what graphics editors /not too complicated to learn, and most of them free, which is quite important, especially for students/. When the picture is too large in size, it is difficult to open and the user often loses interest in your site and leaves without reading your unique content. My task is to show you how to reduce the size of images in some graphics editors.

Let's see how to make an unique image. Immediately I want to warn you - this short review is not intended to teach you how to present other works as your own. We just talk about how to make an image unique to the browsers.

Generally, methods for creating uniqueness are many. Most flexible and reliable is to make yourself a picture, to paint a picture, to make mosaic, painting on stone, fabric or glass, i.e. to do it with your own hands.

However, not everyone has the innate ability to paint, the opportunity to learn to paint or enough time and money to do so.

That's why most people, including myself, use other images, making them unique in using special tools and techniques.

As a tool can be used more or less known graphics editor such as Photoshop, which I will talk about later. If you do not have Photoshop or some other program, try to use the online service [pixlr.com/editor](http://pixlr.com/editor).

In the process of making unique images remember the goal - to ensure that as few copies of the picture as possible will be found via the service [tinEye.com](http://tinEye.com). The best would be if after the inspection the program shows you "zero".

Here are some ways to make a unique image:

First way – Invert

Just tilt the image using a small (or larger) angle to the starting position.

The second way – Reflection

Nothing complicated - mirror the picture horizontally or vertically and then store.

The third way - Cut the sides

You can try to cut all over at once or separately left, right, top, bottom. You can cut diagonals as well if you want.

The fourth way – Collage

In other words, this is the layout of a plurality of images. You find interesting images and combine them into a single and comprehensive picture.

Fifth way – Deflection

Stretch or compress the selected picture vertically, horizontally or diagonally. There are other special types of deformations.

Sixth way - inscriptions and watermarking

Here, I think, everything is very clear. Add words or sentence to the picture or put a watermark - for example containing the domain of your site.

Seventh way – Frame

We put what we need in a framework that we create on our own or find a ready one.

Eight way - Resize / Modifying the size /

Just increase or decrease our (not exactly ours, of course) image. Incredibly, even this simple method is able to give a positive result.

I want to note that not all methods lead to the absolute uniqueness of the photo - you have to try this or that way, to save the image and check again for its uniqueness.

It is advisable to combine several different ways - it helps a lot. You can consider this as an additional way.

So, if you want, you can make each picture unique - try it!

Choose your picture from Google that is larger in size /bytes/ or any of your own photos. Store it as jpeg through the context menu of Mozilla Firefox /I work with this browser/ or using a very handy little program Faststone Capture, we make a screenshot. We use a small rectangle tool to select a part of the picture and save the file. Then change its size by reducing its width to 500 pixels for example. The height is automatically changed, so as to maintain the aspect ratio of the picture. If you set it manually, as we do in <IMG> tag language HTML, we can flatten and deform the image. This, incidentally, may also lead to the uniqueness of the image. Then we can draw something on the picture. Choose your



instrument "Pencil", the thickness and color of the line and draw. You can use text input. We can change the color of the letters, font and font size, thickness and inclination of the letters, and the size of the field in which you write. We can move the text in a convenient location in the image. Finally, store the changed image. During storage of the image press button "Options" and there choose the extension of the image, color palette and image quality that changes in rates by using the slider or Spin Box. Use icon - magnifier with a "+" to bring the image and evaluate visual quality. For WEB we need light images, so leave it for example to 50% and store it. Try again to store the image as a GIF file. Here you can play with colors - instead of 256 we will choose 128 - raster increases. You can choose 64, then the size of the image in bytes is greatly reduced.

Considering other programs allowing us to edit images we present Faststone Image Viewer 5.1. In the Edit menu, choose Resize/Resample. Again, change the size in pixels or percentages. Once you change the width and height changes automatically and then immediately shows you the new amount of megapixels at the top of the command window.

This is a picture of the vault in Washington that my daughter made personally. On the right it shows the original size in pixels and megabytes. We change the width of the picture all the way down three times - from 3648 to 1216. Because we preserve proportions /check in the appropriate box / it automatically reduces the width. We see that the file size is now 2.92 Mb / 28.6 Mb of / - Figure 1.

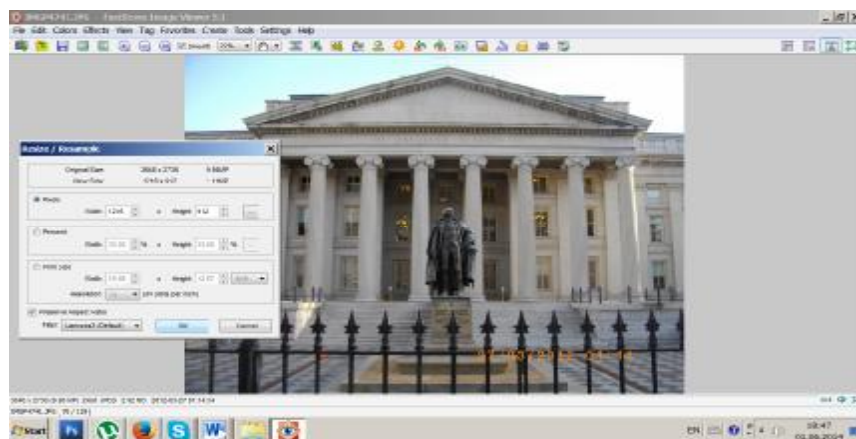


Figure 1

Here, the storage is made also by pushing the button Options, where you can change the quality and filters. Immediately see the

change in size in KB /kilobytes/ of the output image and the finally made image. We use the magnifying glass and monitor the change in the quality of the picture and the change of the file size. If you take a quality of 30%, the file size decreases about 20 times and takes 20 times less time to be uploaded on the site. If you need a better quality, just pick an option in the middle of the chart. /Figure 2/

Figure 2

The program is very easy to work when the image is full screen. Clicking on the four frames on the screen appear different menus that facilitate your work with the image. Figure 3 shows one of them.

Figure 3

There is also a program called Paint.net, the latest version is 3.5.11. There it open windows with the palette of colors, layers and diary which describes the work done so far. We can close them, as it will not work with them now. Copy the image or open the picture file /our case/. Choose the menu Option and submenu Image Resize and then do the resizing of the image. In the program it has a check box to keep the proportions and if active /tick/ the change in one component - width or high, automatically changes the other. We change the width three times, so the hight is changes three times also. We save the resized file by pressing the Save as ... and choose the GIF format. And here's the best part - there is a slider to change the quality /also has a spin box/, and above the image the size is displayed in KB. The panel with the tools is almost like in Photoshop - it has drawing tools, for cropping, adding text, etc. But our goal is not to learn all the possibilities for editing but to see how to change the size in KB in different graphic editors.

Now we open the designers' favorite editor Adobe Photoshop.

Again we will use the photo of the vault in Washington. On the right we see the original size in pixels and megabytes. We change the width of the picture all the way down again three times - from 3648 to 1216. Because we preserve the proportions /check in the appropriate box/ it automatically reduces the width. We see that the file size is now 3.17 Mb /28.6 Mb of/ - Figure 4.



Figure 4

We have opened the image in the program. With the Frame from the Tool box we crop the image and we can change its location /with Shift/. Choose from the menu Image and submenu Image Resize. We

reduce the width and the height is determined automatically, then we store the image in the File menu and submenu Save as Web, and as we do it opens a window that allows the image to be sized. Four options of varying quality appear. From the menu on the right we choose the file extension GIF or JPEG, the quality - again using the slider or spin box. We have to place a mark on Progressive, so to be able to upload the image in parts. If unchecked, it will not open until you fully upload it in the site. The size in KB is shown to the left under the picture /Figure 5/. Save it in our directory.



Figure 5

By Auto button you can choose the best quality, then the size is increased to 43.3 Mb. We are trying to find the optimal option for us /Figure 6/.



Figure 6

When stored in GIF format, if you only change the number of colors in the palette you can observe contraction of the file size. In the following figure you can see top left the original, and then shrinking the file size at 256, 128 and 64 colors /Figure 7/.

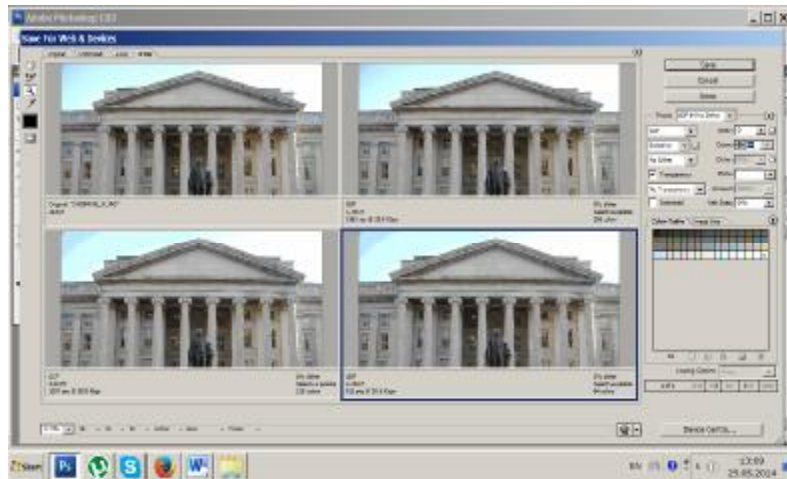


Figure 7

If you choose JPEG format, you get 3 images /save the original/ - low /10%/ medium /30%/ and high /60%/ quality. Below left you can see the file size and upload time in seconds when the upload speed is defined - Figure 8.



Figure 8



If you want, you can also change other parameters such as "blur", colors, text compression /if any/, but it increases the file size.

If we have previously used the function Change the size of the image using the tools of the box, the size would be even smaller.

Using other effects such as rotation, filters, layers and many others from the wide arsenal of Adobe Photoshop will allow us to achieve the uniqueness of the image that we search. Some of these effects were shown in the program Paint.Net, so I will not review them again.

I will show you the results from the check I made with TinEye Reverse Image Search for uniqueness of the output image and this with the added effects /distortion, crest etc. / in the program Paint.Net. Here's the result /figures 9 and 10/:



Figure 9



Figure 10

I hope this report will be very useful in your work, so you will easily manipulate your images and quickly upload them to your site.

#### **References**

1. Help of program Faststone Capture 5.3.
2. Help of program Faststone Image Viewer 5.1.
3. Help of program Paint.net 3.5.11.
4. Help of program Adobe Photoshop CS3 version 10.0.

#### **Author**

**Mariya Deshkova Dimitrova**, Varna Free University, Assistant professor,  
phone +359894366803  
e-mail: [deshkova@abv.bg](mailto:deshkova@abv.bg)

## BIG DATA STORAGE IN HDFS AND HADOOP IN CLOUDS

Vladimir Dimitrov

**Abstract:** *Big data storage is a big problem, because data must be stored on large clusters of thousand servers in a way suitable for it processing. Pioneer in this area is Google, who with its solutions influenced the open source project Hadoop. This paper is an overview on Hadoop Distributed File System (HDFS) and cloud offers of Hadoop .*

**Keywords:** *Big data, Hadoop, HDFS, clouds*

### Introduction

Big data is a collection of data from traditional and digital sources inside and outside the organization that represents a source for ongoing discovery and analysis. They are mix of unstructured and multi-structured data that comprises the volume of information. The global analyst firm Gartner use phrases like “volume” (the amount of data), “velocity” (the speed of information generated and flowing into the enterprise) and “variety” (the kind of data available) to frame the big data discussion. Others have focused on additional V’s, such as big data’s “veracity” and “value.”

The main considerations in big data analytics can be defined by the following questions are:

1. How big data is stored?
2. How big data is analyzed?

Traditional database systems have limitations on the number of columns, table size, etc. Usually, they require data to be preformatted before to load them in the database. Big data, usually, are not formatted. A big data file could store many terabytes and has billions of fields. A special storage for big data is needed.

A solution to above mentioned problems is Hadoop. Usually, when a new technology emerges, there are no standards, but only leading solutions, projects, products, practices etc. Such a leading project in the big data analytics is Hadoop. It combines together data storage and data analytics. HDFS is the Hadoop’s solution of the first problem (Big data storage) and it is discussed here. Hadoop MapReduce is Hadoop’s solution of the second problem.

Apache Hadoop [2] is an open-source Java-based framework. It stores data (Hadoop Distributed File System) and executes jobs (MapReduce) on large clusters of commodity servers. Hadoop supports a high-level of fault tolerance. This framework is very simple but effective for a large class of big data applications. It is scalable from a single server to thousands of servers.

Broadly speaking, Hadoop is an open source implementation of Google File System [3], Google MapReduce [4], and Google BigTable



etc. Google engineers teach, outside the company, on Google technologies using Hadoop.

### **Hadoop Components**

Hadoop core are Common and MapReduce. The current version of MapReduce is MapReduce 2.0 (MRv2) or YARN. All other components are optional. Some of them tend to be implemented in the Hadoop core. HDFS plays special role in Hadoop: it is Hadoop distributed file system, optimized for MapReduce jobs.

MapReduce is a framework for parallel processing of large data sets. It originates from functional programming. MapReduce has very simple programming model, but it is possible useful programs to be written in. There are two phases in MapReduce: Map and Reduce. For every phase, the programmer supply a function with predefined interface. In Map phase, the user-defined function is applied on a key-value pair and generates a list of key-value pairs. Then, in Reduce phase, all lists of key-value pairs are sorted in lists of key – values list, where the values list contains all values generated in Map phase for a specific key. In Reduce phase, user-defined function iterates on key – list pair and generates a list of values. This processing can be pure functional (without side effects) and highly parallel in both phases.

This paper is focused on big data storage in Hadoop - HDFS.

### **HDFS design concepts**

Hadoop flagship file system is Hadoop Distributed File System. This is successor of NDFS [1] (Hadoop predecessor project for web search engine and web crawling on clusters of commodity servers). Hadoop could be linked to any file system that support its abstract file system interface. HDFS is an implementation of Hadoop abstract file system, but it has special place – HDFS is optimized for MapReduce jobs that are core part of Hadoop.

HDFS is suitable for big data analytics. It is optimized to write once very big files and then to read them many times. HDFS is not suitable for random reads/writes. Hadoop complements modern database systems – they are optimized for random I/O operations.

HDFS is a general purpose distributed file system for very large files. Here, under the term “very large data files” must be understood files of many terabytes and petabytes in size. HDFS could store tens of millions files on a single instance. It is designed to run on hundreds or thousands commodity servers. A general assumption about HDFS is that hardware failure is a norm, but not an exception.

HDFS is designed for applications that consecutively access (streaming) the data sets. HDFS is optimized to write once its files and read them many times. It is not suitable for small files and for direct reads

and writes. Applications that run on HDFS use batch processing. HDFS is optimized for high throughput – not for fast access to data. Interactive applications on HDFS are not the standard use case, but such applications could use HDFS under special circumstances. Applications developed in MapReduce framework on HDFS data sets are the primary use case of Hadoop.

HDFS commands are POSIX like. It does not support all POSIX requirements, because its applications does not need them.

HDFS file is a sequence of blocks than are stored in the cluster on multiple servers and on multiple disks. Fault tolerance is supported at block level. HDFS blocks are big ones – 64 MB by default. If a file is smaller than a block it share the block with other small files.

Leading Hadoop idea is to move the computations to the storage nodes but not to move the data to the computing nodes. This approach is sensible when computing programs are relatively small and stored data are big enough. In this case, the network traffic decreases.

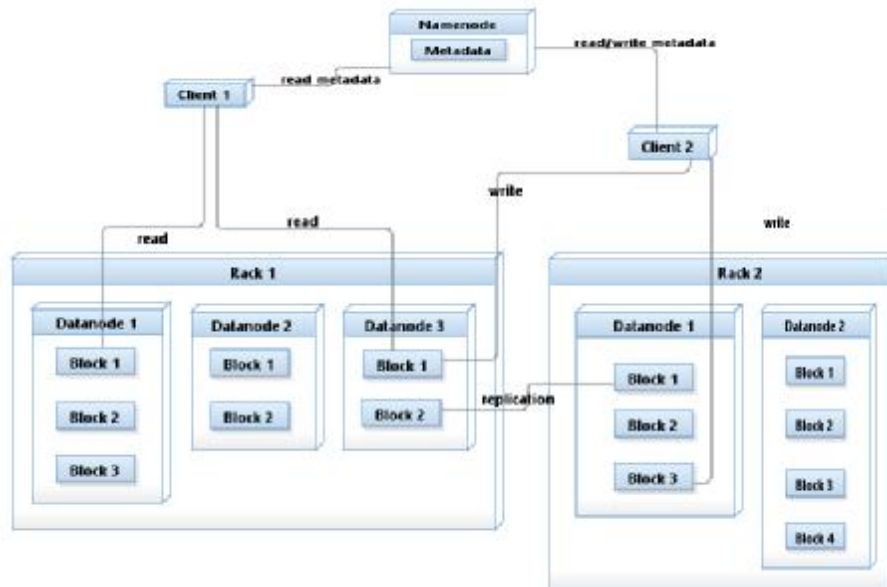
HDFS tries to distribute file blocks among the cluster servers, but at the same time, the size of the file blocks have to be big enough – bigger than the size of the computing programs to decrease network traffic. The maximal parallelism could be achieved if almost all the cluster servers are involved in the computations, but this means that the data set files have to be big enough to be distributed on almost all the cluster servers. Hence, the optimal files for Hadoop are the big one dispersed on large clusters. The file blocks have to be bigger than computing programs and the file blocks has to be evenly distributed on almost all cluster servers – that is the key for Hadoop success.

HDFS is responsible for the even distribution of the file blocks on the cluster servers, but it is not responsible for moving computing programs to the data. The last responsibility is to the MapReduce framework.

#### **HDFS architecture**

HDFS architecture [5] is given at Fig. 1. There are two kind of HDFS nodes: **namenode** and **datanode**. The namenode contains meta-data for the file system: files, directories, file block locations etc. The datanode stores data blocks.

The distance between nodes is defined in the hierarchy: in the same host, in the same rack, in different racks. That metrics is used to calculate when a replica is nearby.



**Figure 1. HDFS Architecture**

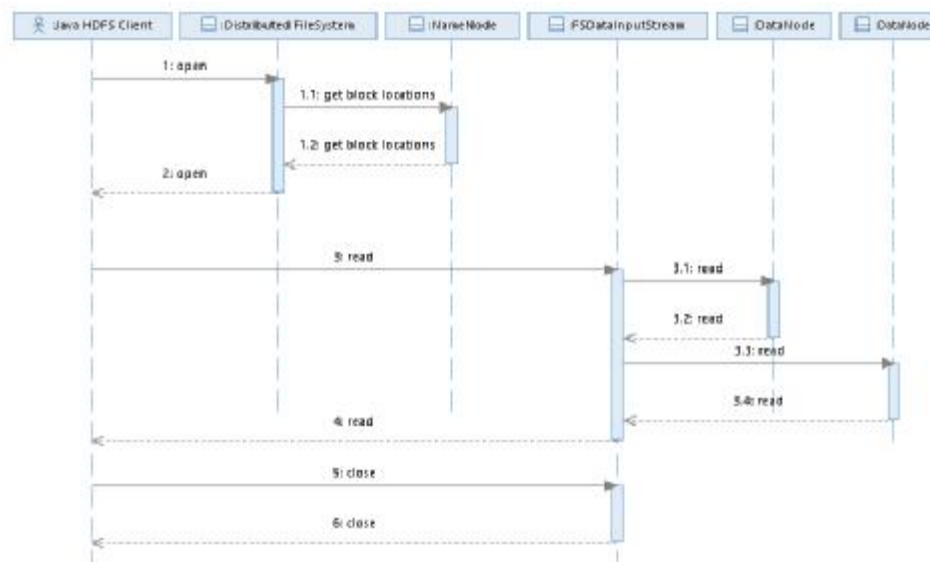
The namenode is crucial for the file system existence – if it fails then all the file system data is lost. That is why there is an option for secondary namenode to be run. The last one collect all the information needed to restart a new namenode in case of failure.

In HDFS version 2.xx, HDFS Federation is introduced. It permits the cluster to be scaled. New namenodes could be added. Every namenode manages different part of the file system – **a namespace volume**. The last one contains metadata information for the namespace and a block pool that contains all block for the files in the namespace. Namespace volumes do not communicate with each other. One datanode can store blocks from multiple namespace volumes. It contains information about that.

In case of namenode failure, a recovery can be done in 30 minutes or more. That is an HDFS problem – the high availability. The good news are that there are plans for future implementations to resolve this problem with a solution based on ZooKeeper. This solution automatically switches to secondary namenode and fences the failed node for all operations that can damage the data.

The client opens files or directories using the metadata from a namenode. After that, all operations are performed directly with the file's datanodes. In case of read operations, datanodes are accessed directly in sequential read access mode. In case of read failure a block replica is

read. The failed block is marked for further recovery operations. It is reported immediately to the namenode before to continue with further reading. When all data from a given block are read, HDFS chooses the next block among all replicas – it must be closer to the client. An example of sequence of read operations in HDFS is given in Fig. 2.



**Figure 2.** Read operations in HDFS

When write operation is performed the block is piped through all file data blocks to the end – write operations are in sequential file access mode. All write operations are queued in the client (FSDataOutputStream). Success of write operation is piped backward to the client and removed from this queue. If a write operation fails then that information is returned to the client, then recover actions are performed to remove failed datanode and to redo all not acknowledged operations. An example of sequence of write operations is given in Fig. 3.

It is possible to put in synchronization points among blocks to start reading from some such an intermediate point.

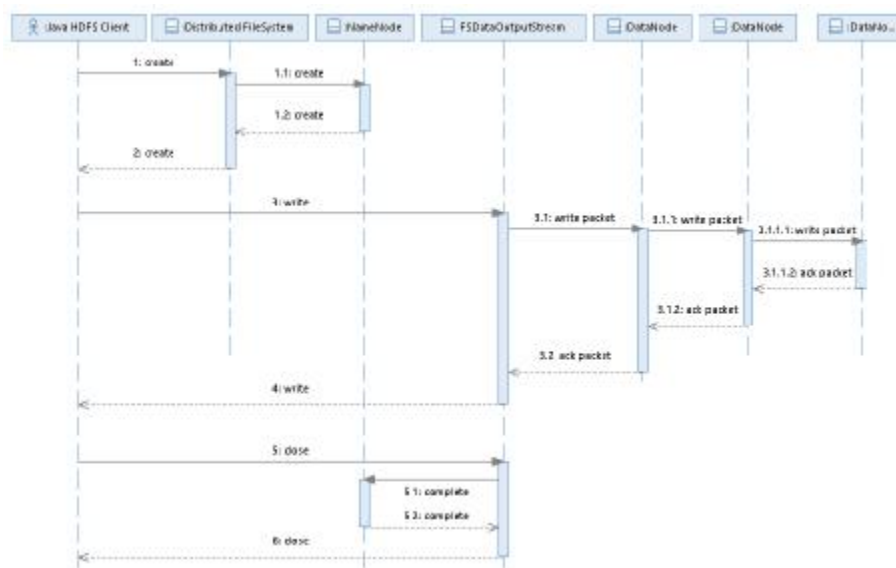
There are many interfaces to HDFS that can be used to access data, in Java, HTTP, C, FUSE etc.

Hadoop support two kinds of file based data structure: SequenceFile and MapFile. The first one is suitable for map inputs. The second one – for reduce inputs. The MapFile is a kind of persistent implementation of Java maps.

There are some utilities for managing HDFS, like Data Ingest for moving large stream data into HDFS and Sqoop for importing structured data stores (relational data bases).

Hadoop archives (HAR files) can be used to package many small files. HAR files are acceptable input for MapReduce programs. They do not support archive compression. HAR archives are immutable, i.e. they cannot be changed – if a file in it is changed then the whole archive has to be recreated.

Hadoop supports file compression, but it is important compression schema to know the block boundaries for splitting. Such schemas are bzip2, and LZO schema with indexing. For big data, the file compression reduces storage needed for the file and speed up data transfers.



**Figure 3.** Write operations in HDFS

Hadoop is a Java implementation. It is important for RPC to support serialization. Hadoop uses its own serialization format Writables. It is compact and fast, but not extensible for other programming languages. Avro is the serialization solution of Hadoop, based on JSON. It is extensible and interoperable.

### Case Studies

There are many Hadoop case studies in different areas, like politics (Obama re-election), financial services, health care, human sciences (NextBio), telecoms (China Telecom Guangdong, Nokia), entertainment

(Orbitz), logistics (US Xpress), energy (Chevron, OPower), retail (Etsy, Sears), data storage (NetApp), software (SalesForce, Ancestry), imaging (SkyBox, Comcast), online publishing (Gravity) etc., but they are reported only in brief. For more details see Hadoop Wiki [6].

All Hadoop use cases show that it is really scalable on large commodity clusters and the simple MapReduce framework is successfully applicable for large class of big data analytics on different media files, logs, indexing etc.

#### **Hadoop on Google Cloud**

Hadoop with HDFS, in principle, originates from Google MapReduce and Google File System. Hadoop could be viewed as an open source of these Google solutions. On the other hand, Google does not offer direct access to its MapReduce. Google stimulates usage of Hadoop on its cloud platform. Detailed information about how to use Hadoop on Google cloud is published on the site of Google Compute Engine. Several scenarios are offered and as options Hive and Pig are regarded. Google offers a connector for Hadoop to its cloud storage, but Hadoop MapReduce is optimized for HDFS. There are sources that evaluate Hadoop as better solution than its Google origins.

#### **Hadoop on Amazon**

Amazon Elastic MapReduce (Amazon EMR) is a Web service that offers Hadoop on Amazon EC2. Amazon EMR could be started in minutes. The service is responsible for Hadoop configuration. Data could be imported/exported from/to Amazon S3. There are detailed instructions and code templates how to develop applications with Hadoop tools like Hive, Pig, and in different programming languages like Java, Ruby, and Perl. Cluster monitoring is available. Running cluster could be easily expanded. When the job is done, the cluster is automatically removed.

#### **Hadoop on Microsoft Azure**

Microsoft offers Windows Azure HDInsight that is 100% Apache Hadoop. It is a service for deployment of Hadoop cluster in the Microsoft cloud. An alternative offer is Hortonworks Data Platform (HDP) on Windows (from Hortonworks) for in-premises cloud deployment. HDInsight is integrated with Microsoft Business Intelligence platform, with relational (MS SQL Server) and non-relational databases via Polybase available with SQL Server 2012 Parallel Data Warehouse, with Tables – NoSQL key-value storage and Blob – a storage for big objects.

With HDInsight can be created Hadoop cluster in minutes on Microsoft Azure and so quickly the cluster can be removed. The Hadoop cluster runs on Linux virtual servers in Windows Azure. There are Hadoop command-line interface and a portal for Hive and Pig.

### **Hadoop on IBM Cloud**

IBM Platform Symphony with IBM InfoSphere BigInsight is the IBM's grid (cloud) offer. IBM InfoSphere BigInsight is a Hadoop implementation of IBM. Many Hadoop components are rewritten for optimization and compatibility. MapReduce is implemented as Adaptive MapReduce that is faster than the original implementation as shown by STACK. Adaptive MapReduce is optimized for adaption to the user needs and the system workload using user defined metrics.

Big SQL is an alternative of HiveQL in this offer. It works not only with big data via MapReduce but directly queries small data.

General Parallel File System (GPFS) is offered as an alternative to HDFS in this solution. It is POSIX compatible distributed file system developed in IBM for big data storage.

Eclipse based tools are available for development of Big data applications.

Console for monitoring and access control is available in IBM InfoSphere BigInsight.

Big Sheets is Web based tool for analyzing and visualization of big data that is part of the offer. IBM InfoSphere BigInsight is integrated with InfoSphere Streams (stream processing), InfoSphere Data Explorer (multi-source multi-type data manager) and Cognos Business Intelligence.

### **Packaging**

As many open source projects Hadoop is freely available but hardly installable – versions, subversions, releases, external libraries etc. require high Hadoop expertise to deploy it. That is why there are several deployments alternatives freely offered by commercial vendors: Cloudera, Microsoft, Google etc. They offer prepackaged version of Hadoop that easy can be installed and deployed on the target environment.

Cloudera is a leading promoter of Hadoop. It is better to check their offers before to start Hadoop deployment. Cloudera is initiator and contributor of many open source projects. Its leading expertise is on Hadoop. Cloudera contributes to HBase, Hive, Pig, Hadoop LZO, HTrace, JCarder, JTrace, Jenkins, MooTools, Record Breaker, and the US FDA Adverse Drug Event System. Clouderians are cofounders of the next Apache open source projects: Avro, Bigtop, Crunch, Flume, Hadoop, Hive, Lucene, MRUnit, Oozie, Sentry (incubating), Sqoop, Whirr, ZooKeeper; and other open source projects: Cloudera Development Kit, Crepo, Hue, Impala, Kitten, ML, Seismic Hadoop.

### **Conclusion**

Hadoop processing power is based on HDFS. The last one distribute evenly data on the cluster. MapReduce tasks are executed on the

data clusters. Hadoop differs from traditional computing where data come to the programs.

The level of concurrent execution depends on the number of data servers. So, HDFS task is crucial for Hadoop job execution. It is possible another distributed file system to be used via the Hadoop abstract distributed file system interface, but the concrete file system must take in account the HDFS functionality.

The host distributed file systems usually are not optimized for Hadoop MapReduce.

### References

1. Cafarella, M., and D. Cutting. 2004. Building Nutch: Open Source Search. ACM Queue, (April): <http://queue.acm.org/detail.cfm?id=988408>.
2. Cutting D., 2012. Intro to Hadoop and MapReduce. <https://www.udacity.com/course/viewer#!/c-ud617/l-306818608/m-312934728>.
3. Dean J., and S. Ghemawat. 2004. MapReduce: Simplified Data Processing on Large Clusters. Google (December): <http://labs.google.com/papers/mapreduce.html>
4. Ghemawat, S., H. Gobioff, and S.-T. Leung. 2003. The Google File System. Google (October): <http://labs.google.com/papers/gfs.html>
5. Hadoop. HDFS Architecture Guide. August 4, 2013. [http://hadoop.apache.org/docs/r1.2.1/hdfs\\_design.html](http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html) (accessed March 11, 2014).
6. Hadoop Wiki. PoweredBy. April 23, 2014. <http://wiki.apache.org/hadoop/PoweredBy> (accessed April 23, 2014)

### Author:

Prof. **Vladimir Dimitrov**, Sofia University



## AN ANALYTIC STUDY ON THE SECURITY ISSUES OF CLOUD COMPUTING

Ananthi Sheshasaayee, K. Geetha

**Abstract:** *The cloud acts as a big black box, nothing inside the cloud is visible to the clients. Cloud computing is a compilation of existing techniques and technologies, packaged within a new infrastructure paradigm that offers improved scalability, elasticity, business agility, faster startup time, reduced management costs, and just-in-time availability of resources. Cloud Computing is a security dreadful and it cannot be handled in usual ways. In that Security issues are the most important part of the cloud facing problem. In this paper the cloud security issues and its life cycle based on the detailed specification of the cloud security issues covered by any proposed security solution have been analyzed.*

**Keywords:** *cloud computing; cloud security; level of security*

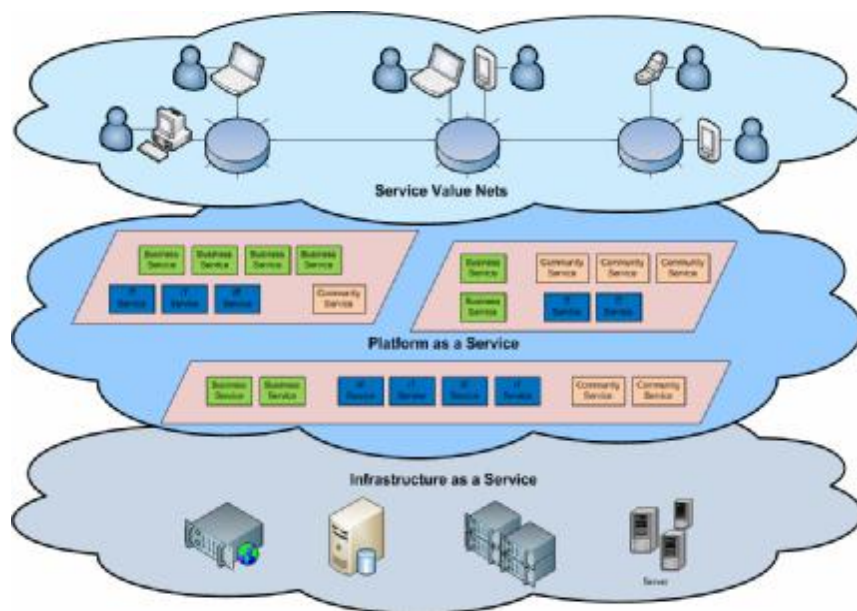
### I. INTRODUCTION

Cloud computing provides the next generation of internet based, highly scalable distributed computing systems in which computational resources are offered as a service. "Cloud Computing", a novel computing model for large-scale application, has attracted much attention from both academic and industry. Since, 2006, Google Inc. Chief Executive Eric Schmidt introduced the term into industry and related research had find-grained management of services through efficient resource sharing mechanism. Convinced by the manner called "pay-as-you-go" [1] [2]. Many cloud services rapidly appear on the Internet, such as Amazon's EC2, Microsoft's Azure, Google's Gmail, Facebook, and so on. However, the well-known definition of the term "Cloud Computing" was first given by Prof.Chellappa in 1997 [3]. Prof. Chellappa recommended that it would be a new "computing where the limitations of computing firm by economic basis rather than technological limit alone". But today is a computing world is cloud computing, it also establishes a kind of business models and merges related technologies together into a single umbrella [4][5][6]. According to the definition of NIST [7], essential characteristics of cloud computing include on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service.

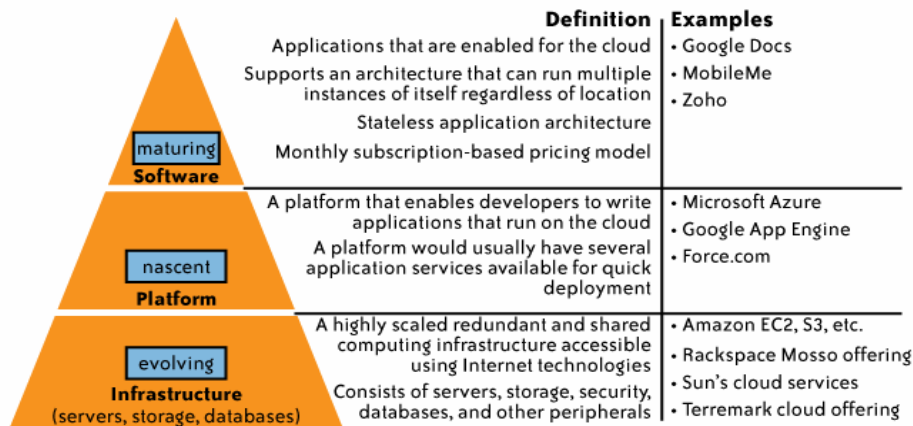
#### A. Cloud Service Models:

Infrastructure as a service (IaaS) provides the consumer with the capability of provision processing, memory allocation, internet, and other resources allow the client to deploy the software, which can include operating systems and applications. The IaaS layer offers storage and infrastructure resources that are needed to deliver the Cloud services. It only comprises of the infrastructure or physical resource. Top IaaS Cloud Computing Companies: Amazon (EC2), Rackspace, GoGrid, Microsoft, Terremark and Google.

Platform as a Service (PaaS) provides the consumer with the capability to deploy onto the cloud infrastructure, client applications, produced using programming languages and tools that supported by the providers. PaaS provides the combination of both, infrastructure and application. Hence, organizations using PaaS don't have to worry for infrastructure nor for services. Top PaaS Cloud Computing Companies: Salesforce.com, Google, Concur Technologies, Ariba, Unisys and Cisco.



Software as a Service (SaaS) provides the consumer with the capability to use the provider's applications running on a cloud infrastructure. The client cannot manage and control the cloud. [8]. In the SaaS layer, the Cloud service provider hosts the software upon their servers. It can be defined as a in model in which applications and softwares are hosted upon the server and made available to customers over a network. Top SaaS Cloud Computing Companies: Amazon Web Services, AppScale, CA Technologies, Engine Yard, Salesforce and Windows Azure.



#### B. Cloud Deployment Model:

**Private Cloud:** The Cloud infrastructure is operated for a private organization. It may be managed by the organization or a third party, and may exist on premise or off premise.

**Public Cloud:** The Cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services.

**Community Cloud:** The Cloud infrastructure is shared by several organizations and supports a specific community that has communal concerns. It maybe exists on premise or off premise.

**Hybrid Cloud:** The cloud infrastructure is a composition of two or more clouds that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability [9].

#### C. AAA (Authentication, Authorization, Auditing):

Three A's are popular concept in security management, in that Authentication process of verifying a whether a user is who is claims to be, and one of the way to verifying the user is password ,but may not be sufficient because lot of password break software are available in the market. In the cloud data is need additional security to maintain the user data.

Authorization is process of making sure that only authorized user can access the data from cloud. Auditing is one of the main security complaint requirement in the cloud computing. Auditing is process of going through all the activity in the system, infrastructure and keeping track of what is happened. In case of an attack or in case of some

anomaly going back end being able to make sure that what happened in the cloud system.

## II. CLOUD COMPUTING SECURITY

### a. Data Leakage:

Data Leakage is a common concern in a traditional IT environment, in case of cloud it become very concern of multi-tenets. There is a concept called side channel Attacks(SCA) where used for data leakage done in cloud by VM which is hosted in the same server can tries to get access of the data from one VM to another VM. The VM may common shared memory.

In the Side Channel Attacks extracts information by monitoring indirect activities. The user in the particular Virtual Machine gets access to the data of another user through common logs or through cache data, then there is a possibility of data leakage. Protection against cross VM Side Channel Attack requires placing only those clients that have no conflicts with one on the same server.

According to many recent surveys, more than 70 percent of IT and business professionals using cloud applications in both personal and enterprise working environments. Data Lack Prevention on the cloud is not a new challenge. However, most of the time, security is the last variable considered, especially the difficulty of tracing and detecting very sensitive data leakage — or worse – the lack of encryption used when storing sensitive data in the cloud. The truth is that nobody is monitoring cloud-based applications.

### b. Denial of service Attacks:

Denial of Service is an example of an attack on availability. It is an attempt to prevent users' data in the cloud. In case of cloud Denial of service becomes more complicated, because enterprise doesn't have to the cloud infrastructure, the cloud providers make sure that all the services and all the resources are available to the user.

Cloud services are becoming increasingly popular these days, both among the public and business enterprises. While convenient, cloud services can be extremely vulnerable to Denial of Service attacks (DoS). A DoS attack makes the network or machine unavailable to the intended users by flooding them with connection requests. Within the eighth annual worldwide infrastructure security report from security provider Arbor Network.

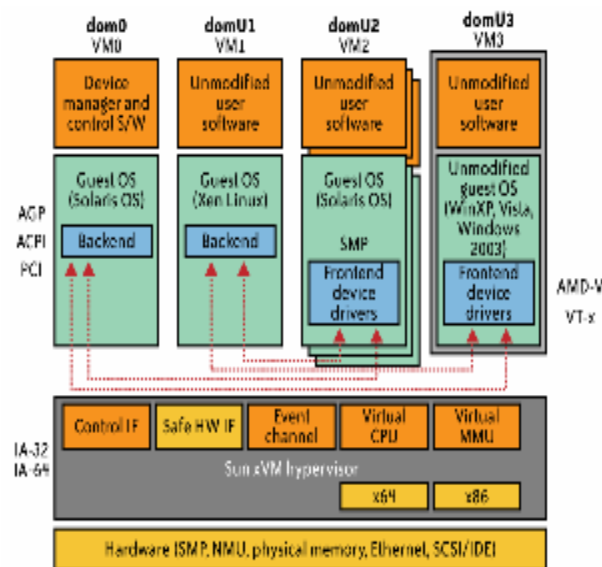
The report indicated 94% of data center operators reported security attacks, 76% had suffered distributed denial of service(DDoS) attacks

towards their customers, while just under half 43% of total infrastructure outages due to DDoS any only 14% respondents seen attacks targeting any form of cloud service.



### c. Host level security

In both the PaaS and SaaS platforms abstract and hide the host OS from end users, Host security responsibilities are transferred to the CSP (Cloud Service Provider, however, as a customer, still own the risk of managing information hosted in the cloud services.



When reviewing hosts security and assessing risks, it is consider the context of cloud services delivery models (SaaS, PaaS, and IaaS) and deployment models (public, private, and hybrid). Although there are no known new threats to hosts that are specific to cloud computing, some virtualization security threats—such as VM escape, system configuration drift, and insider threats by way of weak access control to the hypervisor—carry into the public cloud computing environment. The dynamic nature (elasticity) of cloud computing can bring new operational challenges from a security management perspective. The operational model motivates rapid provisioning and fleeting instances of VMs.

d. Trust:

Trust is not a new research topic in computer science. The concept of trust is adjusted to the case of two parties involved in a transaction, can be described as An user A is considered to trust another user B when user A believes that user B will behave accurately as estimated and required[10].

The trust in a cloud environment depends heavily on the selected deployment model, as governance of data and applications is outsourced and delegated out of the user strict control [11].

### III. CLOUD SECURITY MECHANISMS

In Computer network level security in VM environment all are virtualized in that level how the security is measured and maintained. Some of the mechanism for the cloud security is:

- Computer and network security
- Security data at rest
- Identity and access management
- Risk analysis and compliance

a. Computer level security:

Security at computer level it includes secures the physical server which is part of the physical security and also hypervisor has to be security, so hyper jacking is avoided. Each VM must secure, and security at guest OS level is also very important to secure. Making sure that the OS which is running in that guest OS and also security is maintain at application level. Hypervisor security makes sure that hypervisor is not attack by attackers then. VM Security isolation and hardening is the one of the mechanism to secure the VM. VM isolation helps to prevent a compromised guest Os and applications running on it from impacting other VMs.VM hardening is a process of changing the default configuration in order to achieve greater security.

b. Network level security:

VM is not protected by a normal firewall, and then securing VM to VM traffic running on a server is difficult in a VDC environment. Virtual firewall is a firewall service running on the hypervisor.

i. Demilitarized Zone:

It is a physical or logical network that limits the exposure of the nodes in the internal network from external networks. In the demilitarized zone adds additional layer of security against external attacks. A virtualized DMZ is established in a virtualized environment using virtualized network components.

c. Securing data at rest

It is also secure the data in the cloud which is at rest i.e data which is not, being transferred over a network. If the data are not encrypted raw data stored in the storage then attacker can easily get the data from the cloud. The data is encrypted which is at rest then the providers confidentiality and integrity services protect the rest data in the cloud.

d. Data shredding:

Data which is deleted by a cloud client or a process but which leaves traces on the system can be potential source of attacks. Data shredding is a mechanism to erase the data from the disk permanently and making sure that the deleted data's left form the traces in the cloud.

e. Identity management in cloud

One time password for every new access request new password and measure against "password compromises". Federated identity management is provided as service on cloud to enables organizations to authenticate users of cloud service using the chosen identity provider and also user identities across different organization can be managed together to enable collaboration on cloud.

#### IV. MINIMIZE THE UNSECURE IN THE CLOUD DATA

##### Minimize Lack of Trust

Policy Language: To Create policy language with some of the characteristics are Machine-understandable (or at least process able), Easy to combine/merge and compare. Examples of policy statements are, "requires isolation between VMs", "requires geographical isolation between VMs", "requires physical separation between other communities/tenants that are in the same industry," etc. It is Need a validation tool to check that the policy created in the standard language correctly reflects the policy creator's intentions (i.e. that the policy language is semantically equivalent to the user's intentions).

#### Minimize Loss of Control

Monitoring: Cloud consumer needs situational awareness for critical applications, when underlying components fail, what is the effect of the failure to the mission logic? What recovery measures can be taken (by provider and consumer)? Answer for the above question is requires an application-specific run-time monitoring and management tool for the consumer and the cloud consumer and cloud provider have different views of the system.

Enable both the provider and tenants to monitor the components in the cloud that are under their control to Provide mechanisms that enable the provider to act on attacks can handle.

- infrastructure remapping (create new or move existing fault domains)

- shutting down offending components or targets (and assisting tenants with porting if necessary)

- Repairs

Provide mechanisms that enable the consumer to act on attacks that he can handle (application-level monitoring).

- RAdAC (Risk-adaptable Access Control)

- VM porting with remote attestation of target physical host

- Provide ability to move the user's application to another cloud

Minimize Loss of Control: Utilize Different Clouds The concept of 'Don't put all your eggs in one basket' Consumer may use services from different clouds through an intra-cloud or multi-cloud architecture ,Propose a multi-cloud or intra-cloud architecture in which consumers, Spread the risk, Increase redundancy (per-task or per-application) and Increase chance of mission completion for critical applications

Possible issues to be consider in the utilize:

- Policy incompatibility (combined, what is the overarching policy?)

- Data dependency between clouds

- Differing data semantics across clouds

- Knowing when to utilize the redundancy feature (monitoring technology)

- Is it worth it to spread your sensitive data across multiple clouds?

#### Minimize Loss of Control:

Access Control: Many possible layers of access control to minimize the loss access in the cloud. Example: access to the cloud, access to servers, access to services, access to databases (direct and queries via web services), access to VMs, and access to objects within a VM. Depending on the deployment model used, some of these will be controlled by the provider and others by the consumer. Regardless of



deployment model, provider needs to manage the user authentication and access control procedures (to the cloud).

Consumer-managed access control:

Consumer retains decision-making process to retain some control, requiring less trust of the provider (i.e. PDP is in consumer's domain). It requires the client and provider to have a pre-existing trust relationship, as well as a pre-negotiated standard way of describing resources, users, and access decisions between the cloud provider and consumer. It also needs to be able to guarantee that the provider will uphold the consumer-side's access decisions. It should be at least as secure as the traditional access control model. Face book and Google Apps do this to some degree, but not enough control

## V. CONCLUSION

Cloud computing is sometimes viewed as a reincarnation of the classic mainframe client-server model. However, resources are ubiquitous, scalable, highly virtualized. It Contains all the traditional threats, as well as new ones. Based on the details in this paper summarize the cloud security problems are:

- Some of the security problems are inherited from the used technologies such as virtualization.
- In developing solutions to cloud computing security issues it may be helpful to identify the problems and approaches in terms of Loss of control, Lack of trust and Multi-tenancy problems.
- Security management is very critical to control and manage number of requirements.

## References

- [1] Amazon elastic computes cloud. <http://aws.amazon.com/ec2/>
- [2] Microsoft's windows Azure platform. <http://www.microsoft.com/windowsazure/>
- [3] A webpage of Ramnath. <http://code.google.com/appengine/>
- [4] M. Armbrust, A. Fox, R. Griffith, A.D.Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stocia, and others, "A view of cloud computing", Communications of the ACM, Vol.53, 2010, pp. 50-58.
- [5] I.Foster, Y.Zhao, I.Raicu, and S.Lu, "Cloud computing and grid computing 360-degree compared," Grid Computing Environments Workshop, 2008, GCE'08, 2009, pp.1-10.
- [6] L. M. Vaquero, L.Rodero-Merino, J. Caceres and M. Lindner A. break in the clouds: towards a cloud definition," ACM SIGCOMM computer communication Review, vol.39, 2008, pp. 50-55.
- [7] G. Chockler, et al. "Reliable Distributed Storage," IEEE Computer, pp. 60-67, April 2009.

- [8] D.Zissis, D. Lekkas. "Future Generation Computer Systems," SciVerse ScienceDirect, Dec 2010.
- [9] National Institute of Standards and Technology, the NIST Definition of Cloud Computing. Information Technology Laboratory, 2009.
- [10] International telecommunication union, X-509| ISO/IEC 9594-8. The directory: Public-key and attribute certificate frameworks, ITU X-series, 2001.
- [11] R.Sherman, Distributed systems security, computers&Security 11(1) (1992).
- [12] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona Brandle: Cloud Computing and Emerging IT platforms: Vision, Hype and Reality for Delivering Computing as the 5<sup>th</sup> Utility.
- [13] K. S Suresh, Prof. K. V. Prasad: "Security Issues and Security Algorithms in Cloud Computing. IJARCSSE, Volume 2, Issue 10, October 2012.
- [14] Guide to Computer Network Security, 2nd Edition, Springer, 2013. Joseph M. Kizza. "Cloud Computing Security, Reliability and Availability Issues".
- [15] Cloud Computing: Principles and Paradigms, Editors: Rajkumar Buyya, James Broberg, Andrzej M. Goscinski, Wile, 2011.

**Authors:**

**Dr.Mrs.Ananthi Sheshasaayee<sup>1</sup>**

Associate Professor & HOD, PG & Research Dept of C.S & Application,  
Quaid-E-Millath College, Chennai, Tamil Nadu, India.  
[Ananthi.research@gmail.com](mailto:Ananthi.research@gmail.com)

**K.Geetha<sup>2</sup>**

Research Scholar (Part-Time), PG& Research Department of Computer  
Science & Application, Quaid-E-Millath College, Chennai, Tamil Nadu, India.  
[geethak.research@gmail.com](mailto:geethak.research@gmail.com)

## REDUCTION OF LEARNING SAMPLE SIZE BY FILTERING WELLS ON CORRELATION COEFFICIENT OF APPARENT RESISTANCE

P. Gricenko, Y.I. Kuchin, R.I. Muhamedyev,  
Zh. Nurushev, K. Yakunin

**Abstract:** *Uranium well log data interpretation process can be performed using machine learning systems such as artificial neural network. Such systems create opportunity to get on a more qualitative level of GIS data interpretation, by replacing geophysics interpreter engaged in analysis of huge volumes of similar geological and geophysical information. Key condition for effective usage of such systems lies in high-quality of learning samples. The higher quality of them, the better result of interpretation can be achieved. An approach to filter data to be used for learning sample formation is proposed here. It is based on wells filtering using geographical distance and correlation coefficient of apparent resistance and allows to reduce size of learning sample, as well as to increase quality of recognition.*

**Keywords:** *filtering, size reduction, correlation coefficient, learning samples, uranium well, recognition process, apparent resistivity, geophysics interpreter, neural network.*

### General

For process of uranium extraction using underground leaching method, preliminary geophysical research is vital. During the research electromagnetic properties of rocks are evaluated and interpreted. Results of the interpretation are then used in extraction process.

Data obtained in the course of the geophysical research can be interpreted by a specialist or an intelligent system based on machine learning methods. Interpretation using machine learning algorithms was successfully performed for different types of geological research data [1, 2, 3]. Machine-learning techniques have been studied in detail. In some cases it's possible to achieve results, which are close to optimal, on a given set of data. However, there are several problems, such as preprocessing of data and selecting relevant features for objects to be recognized. Another crucial issue is formation of training samples for different groups of data. In the case of geological research, properties of rocks from different geographical areas can have different properties, which means that creating a single universal training sample and to train a single universal classifier, which would recognize any type of rocks with equal accuracy is practically not possible. The problem of formation of training sample for geophysical research data of wells from uranium deposits of Kazakhstan is covered in the paper.

There are several crucial moments in forming learning sample:

- Data volume;
- Data compliance.

*Data volume.* Uranium deposit can include hundreds or thousands of wells. Moreover, each well can be composed by thousands of well log values. Obviously, it is impossible to process such volume of data, because time to train neural network has direct dependence on learning sample volume. Therefore, only subset of geographically nearest wells should be used to create learning sample.

*Data compliance.* Lithology has layered structure. In different points of uranium deposit layers may have different parameters, like thickness, however if points are close it is reasonable to suggest that lithological structure is approximately the same. Taking wells with lithological structure close to structure of well to be recognized can increase recognition quality.

Well lithological structure can be described by curves (well logs). Similarity of curves can be defined by correlation coefficient. Thus, filtering wells on the basis of correlation coefficient of one of the basic parameters is used to select wells with lithological structure close to structure of well to be recognized.

The major purpose of the study is an experimental evidence of the suggestion that filtering wells based on geographical distance and correlation coefficient of one of the basic parameters can increase quality of recognition.

Actually, the study was based on 38 wells, which contained approximately 37000 dots and 7 classes (Table 1.) 5 wells had been chosen to be recognized. Apparent resistance was used for correlation coefficient calculations. Apart from apparent resistance, spontaneous potential, induction log and depth were used as parameters. Also, 'flowing window' (radius=5, i.e. 5 dots before and 5 dots after current dot) and range normalization[-1;1] were applied to each parameter. Multilayer Neural Network was used as classifier.

TABLE 1 Class distribution of data

<i>Class(Lithotype)</i>	<i>Number of dots</i>	<i>% of whole number of dots</i>
<b>1</b>	13596	36.56
<b>3</b>	13583	36.52
<b>4</b>	4518	12.15
<b>7</b>	2738	7.36
<b>6</b>	2557	6.88
<b>5</b>	187	0.50
<b>9</b>	5	0.013
<b>Sum</b>	37184	100

TABLE 2 Results of recognition of wells lithological structure by neural network

Well	Without filtering		Filtering based on correlation coefficient		Filtering based on geographical distance	
	Points number	Percent, %	Points number	Percent, %	Points number	Percent, %
2092	36189	50,57	4816	48,69	4649	51,62
2093	36201	53,22	5280	56,40	4661	53,44
2253	36257	59,62	4608	55,90	5030	65,58
2254	36291	35,91	5053	46,66	5064	30,06
2257	36006	39,38	4113	44,30	4779	39,39
Average		47,74243427		50,39327508		48,01650904

### Conclusion

According to the results (Table 2), there is an upright trend in average percent i.e. using filtered learning sample gives more quality of recognition, than using learning sample formed by 37 wells (all wells in deposit without well to check). Moreover, volume of filtered learning sample is approximately 7 times smaller.

Additional study should be conducted to check trend on all of 38 wells and on different uranium deposit.

This work is performed by supporting grant №2318/GF3 of the Ministry of Education and Science of the Republic of Kazakhstan.

### References

- [1] Ravil I. Muhamediyev, Yan Kuchin, Elena Muhamedyeva (2012). Recognition of Geological Rocks At the Bedded-infiltration Uranium Fields by Using Neural Networks. 2012 IEEE Conference on Open Systems, Kuala Lumpur, IEEE Xplore database, <http://cs.ieceemalaysia.org/ICOS2012/> 6 p.(accepted 08.2012) ISBN: 978-1-4673-1044-4, Digital Object Identifier : 10.1109/ICOS.2012.6417622M. Tim Jones. Artificial Intelligence: A Systems Approach. INFINITY SCIENCE PRESS LLC Hingham, Massachusetts, New Delhi, 2008. ISBN: 978-0-9778582-3-1
- [2] Kostikov D.V. Tools of interpretation of geophysical researches of boreholes on the basis of the transformed logging diagrams by means of a multilayered neural network.-M.: Russian state library, 2007.- 189 p.).
- [3] Kuchin Y., Muhamedyev R., Muhamedyeva L. Interpretation of log data of boreholes. The 9th INTERNATIONAL CONFERENCE: INFORMATION

TECHNOLOGIES AND MANAGEMENT 2011, April 14-15, 2011, Information  
Systems Management Institute, Riga, Latvia, ISSN 1691-2489, 98 c.

**Authors:**

**R.I. Muhamedyev**, professor, **Y.I., Kuchin**, m.c.s,

**P. Gricenko, Zh. Nurushev, K. Yakunin**

International IT University, Department of Computer Science, Software  
Engineering and Telecommunication

Manas/Jandosov Str., 34A/8A Almaty, Kazakhstan,

e-mail: [ravil.muhamedyev@gmail.com](mailto:ravil.muhamedyev@gmail.com)

phone: +7 707 6844370,+7 777 2399788

## USING NODE.JS IN DEVELOPING WEB SERVICES

Krasimir Ivanov, Evgeniya Rakitina

**Abstract:** *Node.js is a software platform for scalable server-side and networking applications. Node.js applications are written in JavaScript, and can be run within the Node.js run-time on Mac OS X, Windows and Linux with no changes. Node.js applications are designed to maximize throughput and efficiency, using non-blocking I/O and asynchronous events. Node.js applications run single-threaded, although Node.js uses multiple threads for file and network events. Node.js is commonly used for real time applications due to its asynchronous nature. Node.js internally uses the Google V8 JavaScript engine to execute code, and a large percentage of the basic modules are written in JavaScript. Node.js contains a built-in asynchronous i/o library for file, socket and HTTP communication. The HTTP and socket support allows Node.js to act as a web server without additional web server software such as Apache.*

**Keywords:** *Web services, Node.js, non-blocking I/O, JavaScript, V8 engine, asynchronous i/o library, HTTP server*

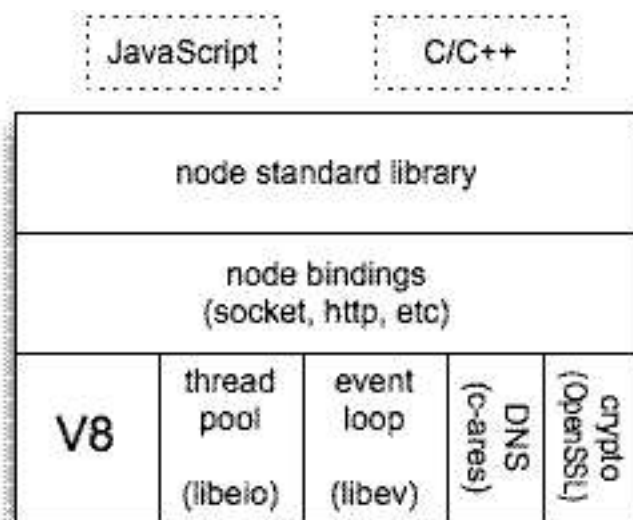
Goal: Introduction to the new technology for developing web services and applications that are designed to maximize throughput and efficiency.

The main problems faced by developers of web services today is scalability of applications the build. Traditional web servers in many cases can't manage to work with enough simultaneous connections and deliver the result in reasonable time. Traditional web-serving techniques require each connection (request) creates a new thread, taking up system RAM and eventually maxing-out at the amount of RAM available. Node.js is different. It operates on a single-thread, using non-blocking I/O calls, allowing it to support tens of thousands of concurrent connections, without worrying about RAM limitations and the cost of context-switching between threads. The design of sharing a single-thread between all the requests means it's good at building highly concurrent applications. The design goal of a Node application is that any function performing an I/O must use a callback. Node.js brings a new element to the mix in the form of an event-driven focus on the server-side. Client-side operational methods simply aren't efficient enough for the current and forthcoming generations of web systems. Node.js pushes data to clients if and only when appropriate. [R1]

Node.js is a platform built around Google Chrome V8 Javascript engine, to create lightweight, fast, scalable, event-driven and non-blocking I/O applications. V8 is an open source project by Google and it is in the very core of Google Chrome browser. Its first public release was on September 2, 2008, same date Chrome's first release was announced. What make V8 so fast is that it's written in C++ and its biggest revolution was that it precompiled the Javascript source code to

machine code instead of just interpret it and then applied a JIT<sup>1</sup> process again in run time to improve dynamic code execution. Node.js was initially released on 31.05.2009 by Ryan Dahl. On 30.09.2009 it is released NPM<sup>2</sup> which helps developers to install Node.js applications on server and manages internal dependencies. On 02.02.2010 starts site “howtonode.org” it is devoted to working with Node.js. On 08.07.2011 Node.js is introduced as project in GitHub. Since then Node.js is developing fast and it has many users.

Traditional web servers do blocking operations for example when there is request for large amount of data from database the server stops and wait for database to respond and then send the data to client. In tradition servers this situation has some resolution on operating system level by using threads and for each request server starts a new thread that leads to allocating a certain amount of memory. In this approach, the main drawback is the relatively small possible number of simultaneous threads and relatively large amount of needed memory. Another method for avoiding the problem of the delays is to use asynchronous programming. However, it is more complicated and in some cases leads to “race conditions” when more then one thread try to access same resource.



*Illustration 1: Node.js architecture*

<sup>1</sup> Just-in-Time Compiler (JIT compiler)

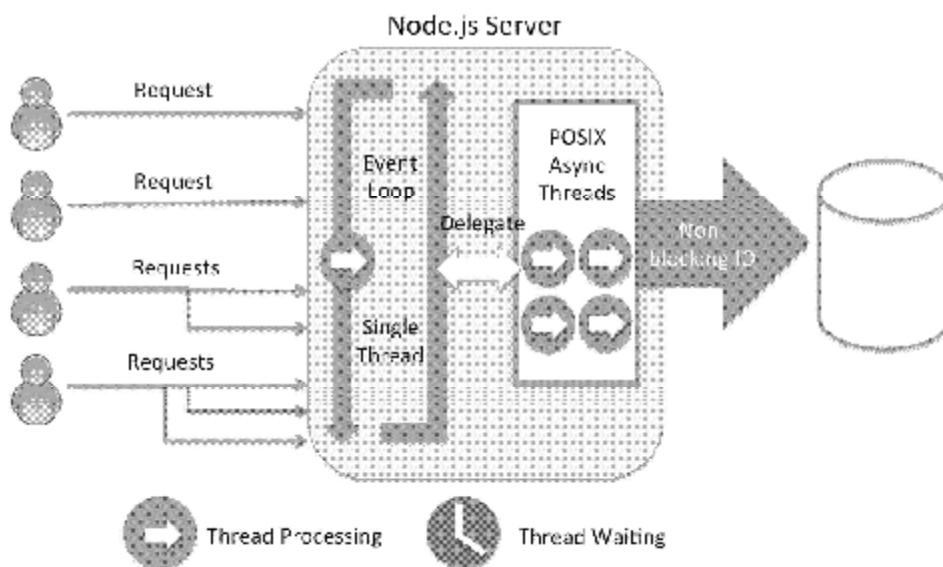
<sup>2</sup> Node Package Manager



Node.js uses the idea to implement callback requests. This means that when in code there is defined function its execution is requested, but not in that moment. The requested function is send for execution and Node is ready to take next request. Later when requested function is executed and return result the result is return back to client. Node.js is written in Javascript, it gives full control over the server. Node.js work in single-thread and work synchronous, but its modules for file system, database and others works asynchronous. This means that the event loop is running and it is always ready to take new request or deliver result returned from finished function. This gives to Node.js ability to manage large amounts of concurrent requests.

Node.js is based on the following technology.

- Libuv - high-performance event driven I/O library - used for reading and writing to the hard disk. It is written in c++.
- V8 - Google Chrome's JavaScript engine.
- JavaScript - calling JavaScript functions leads to call c++ functions that give to server great speed.



#### Development tools:

Only needed tools are text editor and console. However there are IDE that support node like:

- JetBrains WebStorm/PhpStorm (most comprehensive);
- Komodo IDE (full Node.js debugging support since version 8);
- Nodeclipse, an Eclipse plug-in for Node.js

- Cloud9 – cloud-based IDE
- Aptana Studio
- vim
- Sublime Text
- Nide

#### **Steps for installation:**

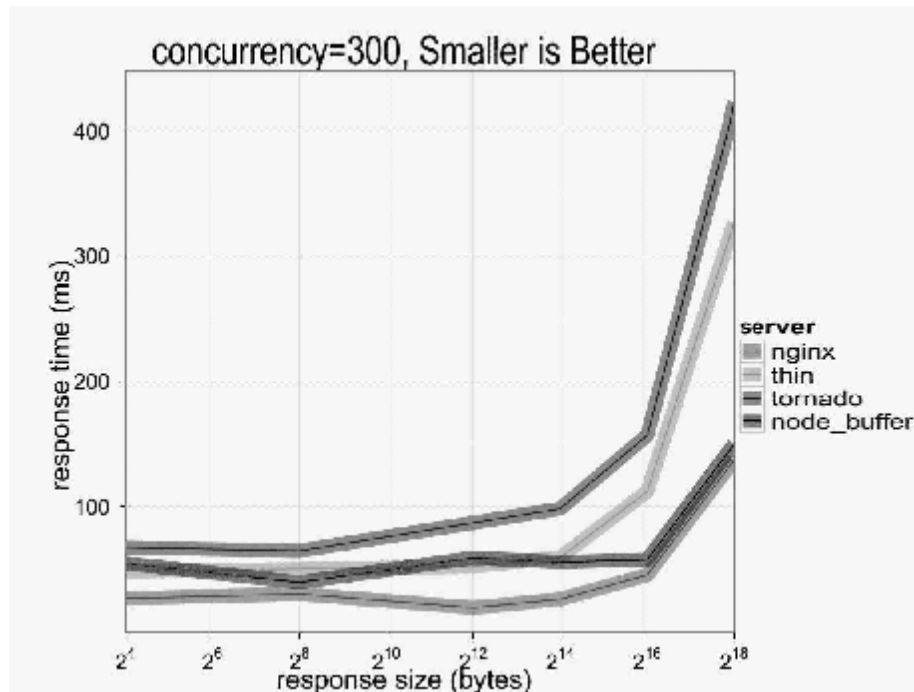
1. Installing npm
2. Installing Packages
3. List all packages: npm ls
4. Updating packages: npm update
5. Development: npm link
6. Making a Package: The package.json file.
7. Publishing: npm publish

In the presence of so many different server technology java script differs in that the language is easy to learn and is more flexible than other programming languages used for server applications.

#### **Example of simple http server written in node.[R2]**

```
var http = require('http');
http.createServer(function (req, res) {
  res.writeHead(200, {'Content-Type': 'text/plain'});
  res.end('Hello World\n');
}).listen(1337, "127.0.0.1");
console.log('Server running at http://127.0.0.1:1337/');
Example connecting Node.js to MySQL:
```

```
var http = require('http'),
    var mysql = require("mysql");
var connection = mysql.createConnection({
  user: "root",
  password: "",
  database: "db_name"
});
http.createServer(function (request, response) {
  request.on('end', function () {
    connection.query('SELECT * FROM your_table;', function (error, rows, fields) {
      response.writeHead(200, {
        'Content-Type': 'x-application/json'
      });
      response.end(JSON.stringify(rows));
    });
  });
}).listen(8080);
```



Example of SOAP server:

```
var soap = require('soap-server');
function MyTestService(){
}
MyTestService.prototype.test1 = function(myArg1, myArg2){
  return myArg1 + myArg2;
};
var soapServer = new soap.S SoapServer();
var soapService = soapServer.addService('testService', new MyTestService());
soapServer.listen(1337, '127.0.0.1');
```

Some of the most popular NPM modules today are:

- **express** – Express.js, a Sinatra-inspired web development framework for Node.js, and the de-facto standard for the majority of Node.js applications out there today.
- **Connect** – Connect is an extensible HTTP server framework for Node.js, providing a collection of high performance “plugins” known as middleware; serves as a base foundation for Express.
- **socket.io** and **sockjs** – Server-side component of the two most common websockets components out there today.
- **Jade** – One of the popular templating engines, inspired by HAML, a default in Express.js.

- mongo and mongojs – MongoDB wrappers to provide the API for MongoDB object databases in Node.js.
- Redis – Redis client library.
- Coffee-script – CoffeeScript compiler that allows developers to write their Node.js programs using Coffee.
- underscore (lodash, lazy) - The most popular utility library in JavaScript, packaged to be used with Node.js, as well as its two counterparts, which promise better performance by taking a slightly different implementation approach.
- Forever – Probably the most common utility for ensuring that a given node script runs continuously. Keeps your Node.js process up in production in the face of any unexpected failures.

Above mentioned advantages of this technology makes it very good choice for developing web services. The strength of Node.js is in serving extremely fast big number of simultaneous connections, that are not computational heavy, but I/O intensive. Node.js shows that can perform very well when building for example REST server, RSS server, Ajax server. Node.js shows that can serve static files very fast, as shown on diagram before node.js serve 256kb files at 300 concurrent connections for about 150ms which is almost good as nginx. Node.js also shows its power when it is used for example in auto completing forms because there are many concurrent connections but low CPU usage, and the data for auto-completing can be stored in NOSQL data base as MongoDB and other non relational databases. Node.js also provide ways to scalability first multiple nodes can be run on multiprocessor systems. Second if there is used proxy server like RabbitMQ which receive connections and redirect queries to multiple node.js instances this method provide easy scalability without changes in application code, and if given node.js instance fails other takes the load.

### References

1. <http://en.wikipedia.org/wiki/Nodejs>
2. <http://code.tutsplus.com/tutorials/nodejs-for-beginners--net-26314>
3. Colin Ihrig, Pro Node.js for Developers, Apress, 2013, ISBN: 978-1-4302-5860-5
4. Pedro Teixeira, Professional Node.js: Building Javascript Based Scalable Software, Wrox, 2012, ISBN: 978-1-1181-8546-9

### Authors:

**Krasimir Todorov Ivanov**, Assistant engineer, Varna Free University, 0887380054, [krasimir19@yahoo.com](mailto:krasimir19@yahoo.com)

**Evgeniya Rakitina**, PhD, Associate professor, Varna Free University, 0887498358, [rakitina@vfu.bg](mailto:rakitina@vfu.bg)

## APPROACHES TO VISUALIZATION OF APOPTOSIS MODELS

A. Mansharipova, E. Muhamedyeva, R. Muhamedyev

**Abstract:** Apoptosis or programmed cell death plays an important role in many physiological states and diseases. Detection of apoptotic cells, tracing the development of apoptosis, drug development and regulation of apoptosis are an important parts of basic researches in medicine.

*This paper discusses the simulation and visual presentation models of apoptosis. Modeling languages, such as SBML, FieldML and CellML are briefly described. The medical literature that shows the possibilities of SBML CellML in apoptosis modeling is analyzed. There is a description of the simulation environments and software libraries for these languages (SBMLeditor, libSBML, SBMLToolbox, MGSmodeller, OpenCell). The approaches to creation of apoptosis visualization system are described on the basis of modeling languages, as well as using of descriptive (qualitative) process models.*

**Keywords:** visualization, chemical kinetics, cell biology, apoptosis

### Introduction

The visualization of biological data is the part of bioinformatics that concentrates on issues of applying the computer graphics in various fields of life sciences. Only a partial list of problems is to be solved in this area including the visualization of the genome, macromolecular structures, phylogeny, dentition, microscopy, and magnetic resonance visualization data. Software visualization systems are also diverse, ranging from simple and straightforward to complex integrated software solutions [1]. Visualization is widely used for educational purposes. However, different methods of modeling are used for the disclosure of essence of the processes, the laws of functioning and the description of systems. The modeling of biological systems is one of the most important tasks of system biology and bioinformatics. The computational biology develops algorithms, data structures and methods for visualization and modeling biological systems. The methods of computer modeling and appropriate modeling languages that are based on XML are often used in solving problems. Generally accepted standards are SBML [2], BioPax and CellML [3]. There are also new projects, which include FieldML [4,5] - the language that describes the spaces (fields), which is used to describe the human body.

This work is a brief and apparently not complete overview of languages and environments for modeling of biological processes that are used and could be used to describe apoptosis. There is also an analysis of the approaches to the creation of system of visualization of apoptosis models.

The work consists of three parts.

The first part is a brief overview of the methods and models of apoptosis on the basis of SBML and CellML.

The second discusses the modeling environment to ensure the application of the languages mentioned above.

The third part analyzes the approaches to the creation of visualization system of apoptosis models. Three basic ways of visualization are described:

1. The creation of animated (video clips) or static visualization of the process by using descriptive (qualitative) model.

2. Establish a system for converting descriptive models with a limited set of parameters in a visual model.

3. Develop a visualization system for converting the changes of parameters in a model that is created by one of the modeling languages.

Note that this review does not address to ways of apoptosis visualization of living organisms by means of invasive and non-invasive methods, which are another big and promising field for future researches.

### **The modeling of apoptosis processes based on SBML and CellML**

The SBML language is focused on representation of chemical reactions in the form of processes. The processes in this case are caused by the actions of participants. Such approach is typical also for the other types of processes. Moreover, the language supports mathematical functions and discrete processes, thus it can be used not only for the modeling of biochemical processes.

One of the models of apoptosis process on the basis of SBML is given in [6], where MatLab is used as the simulation tool.

CellML – is XML-based markup language for describing mathematical models. It is originally created for biology. CellML is currently popular as a compact format to describe the computational models. Although the language reminds SBML, though it is aimed at a wider class of models, not only in the field of biochemistry.

Currently, there are a significant number of sources that use models constructed on the basis of CellML for the analysis of the impact of various factors on the process of apoptosis. For example, in work [9] the authors propose a mathematical model for mitochondria-dependent apoptosis, in which kinetic cooperativity in the formation of apoptosome is a key element to ensure bistability. In the article [10] authors presented a mathematical model which includes TNF initiated survival and apoptotic cascades, as well as nuclear transcription of I $\kappa$ B. Models based on CellML are given in [7], etc.

## Overview of modeling environments

SBMLEditor - relatively simple, low-level editor for SBML files. It allows manipulating the elements of SBML and save the file with the actions. SBMLEditor also allows creating and editing annotations, in accordance with the SBML specifications. [14 11]

libSBML – is a free programming library, with open source reading, writing, manipulation, transfer and checking SBML files and data streams. This is an API (application programming interface) which can be inserted into your own applications. It is supported by many programming languages and operating systems. Referred to C + +, Java, Python, MATLAB, C #, Ruby, Perl, Octave [12, 13].

SBMLToolbox.- is built on the basis of libSBML and provides a set of basic functions that allow using SBML models in MATLAB and Octave. SBMLToolbox provides functions for creating and testing models, manipulating and modeling these models by using ordinary differential equations solvers. SBMLToolbox translates SBML models to / from the structure of MATLAB, called MATLAB\_SBML [14]. It provides tools for working with these organizations as part of the MATLAB or Octave. LibSBML allows importing and exporting these structures to / from SBML files. The toolbar should not be considered complete platform of biology tools for MATLAB, it provides the ability to access SBML from MATLAB and actually serves as a starting point from which users can develop their own applications.

SBML converters – there are a lot of converters that allow converting from/to SBML to BioPax 2 and 3, XPP and Octave. [2]

MGSmodeller - is a system designed for the creation, calculation and analysis of molecular genetic systems (IGU). It contains the following modules: "designer / modeler", "model calculations", "inverse problem", "optimal control". The versatility of the system is based on the capabilities of each module. The module "designer / modeler" allows to create and edit models by using the original and developed by the authors specification standard for «SiBML» models. Syntactic and semantic analyses for language modeling are implemented in this system. Mathematical models in MGSmodeller are reconstructed in the format and rules of the standard SibML within the universal chemical-kinetic approach [15, 16 ]. The analysis of simulation results are produced by MGSmodeller, Matlab and Gnuplot. Compilation units and numerical study is implemented in FORTRAN. Modules annotations and editing SibML language, as well as post-processing of the results are realized in Java language [17]. MGSmodeller system also includes: tools for constructing IGU models, tools of preparing models for the calculation of dynamics, the solution of reverse problems and optimal control problems.

OpenCell (Physiome) - is a framework for working with CellML models. It provides a uniform way of working with CellML documents, including the creation of new models, editing of existing models, and running simulations [18]. Main Features:

- Integration of CellML models that include ordinary differential equations.
- CellML 1.0 and CellML 1.1 support, which allows building complex models by importing one model to another.
- Ability to work in a variety of operating systems, Linux, Windows and Mac OS.
- Ability to build and edit graphics.
- Support of draft specifications metadata modeling and graphics.
- The project is open (Open Source).

The combination of modeling and visualization makes it able to create an appropriate visual representation of occurring processes and explore the impact on them. Thus, one of the objectives of prospective studies can be creating of visualized micro and macro models mechanisms of apoptosis.

There are three potential approaches for apoptosis visualization:

1. The creation of animated (video clips) or static visualization of the process by using descriptive (qualitative) model.
2. Establish a system for converting descriptive models with a limited set of parameters in a visual model (picture, photography, drawing, and video) from a limited array of models.
3. Develop a system for converting the changes of parameters in a model that is created by one of the modeling languages, such as CellML, to changes in visual object model.

The brief analyzing of the capabilities of each approach is shown below.

#### **The creation of animated (video clips) or static visualization of the process by using descriptive (qualitative) model**

This direction was realized during the first phase of the project [22 19]. The following videos were developed:

1. Apoptosis & necrosis.
2. Myocardial tissues of rat with coronary artery disease. Video clip created from photos.
3. The visualization of anti-apoptotic hsp70 warm shock protein function scheme.
4. Animated scheme of apoptosis.



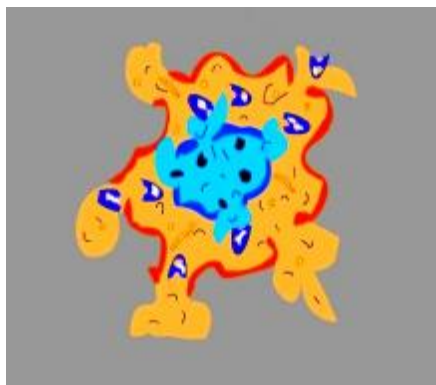


Figure 1. Apoptosis via necrosis.  
Comparative video



Figure 2. Myocardial tissues of rat  
with coronary artery disease

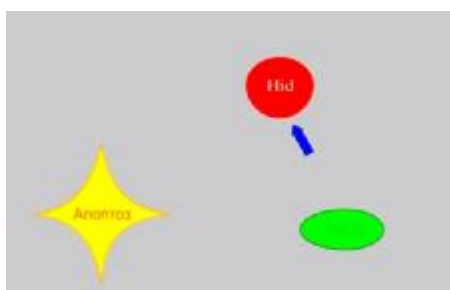


Figure 3. The visualization of  
anti-apoptotic hsp70 warm shock  
protein function scheme

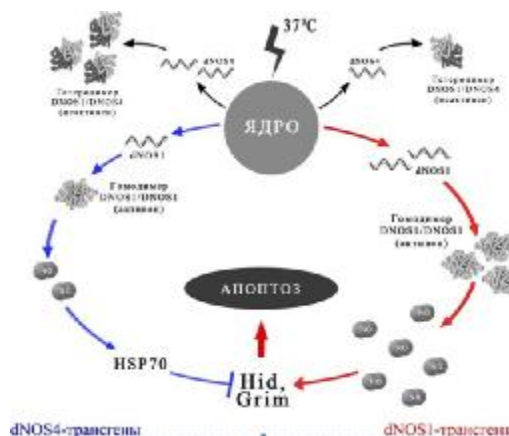


Figure 4. Animated scheme  
of apoptosis

This approach has merits and demerits:

**Advantage** – the simplest approach in realization.

**Disadvantage** – absence of variability, disability of process modeling.

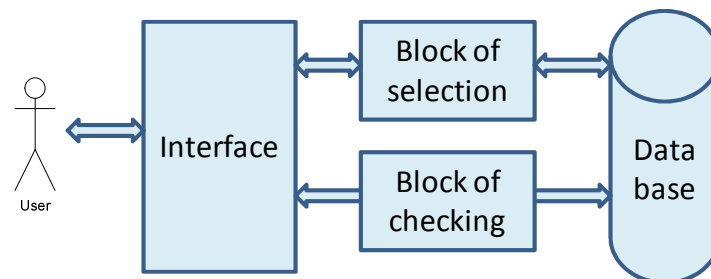
**The convention of descriptive models with a limited set of parameters in a visual model (picture, photography, drawing, and video) from a limited array of model**

During this approach a set of video clips or images (display images) are created, each of which corresponds to its set of parameters. Interface that allows inputting parameters that are limited with scope is created. In virtue of input parameters, the system selects the images that are closest to the input parameters. This approach is possible due to the fact that the available models are descriptive, and if it is planned to use them mainly for educational purposes. In addition, it is possible to create a system to select the most appropriate representation as described in accordance with by means of input parameters. There are several aspects to solve the problem:

1. Create a set of visual images by using a database with a description of the process parameters that is visualized by the image.
2. Create a system that selects the image with the suitable input parameters.
3. Create an interface system, which should analyze input parameters, that prohibits entering inconsistent values and check the validity of data thereby display the result.

The system is illustrated by Fig 5, where User – system user (player, student, doctor); Interface – block that works with the user, implements receiving of model parameters and its visualization; Block of selection – selects the model suitable for parameters input by the user; Database – system database that saves a set of images and its parameters, parameters scope, inconsistent values.

Block of checking – block that controls input parameters.



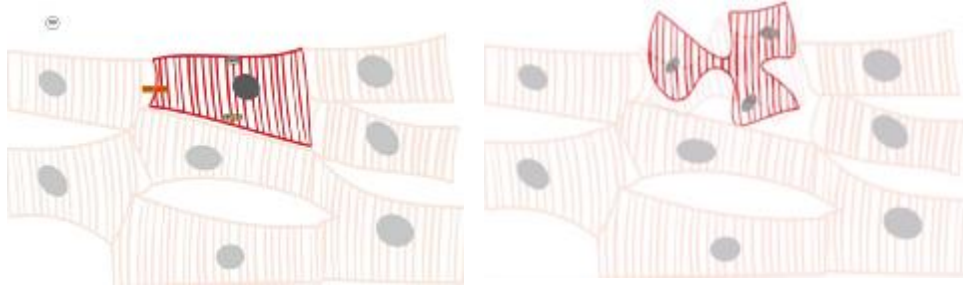
*Figure 5. Interface visualization system*

As above, the method has advantages and disadvantages:

**Advantage** – the simplest approach in realization.

**Disadvantage** – absence of variability, the abilities of process modeling are also reduced.

This approach was realized by creating the simple game based on the model of bistability of apoptosis. The game was created by using game engine - Game Maker. Program realizes the model described in [9]. The figures 6 and 7 show the stages of apoptosis in the game.

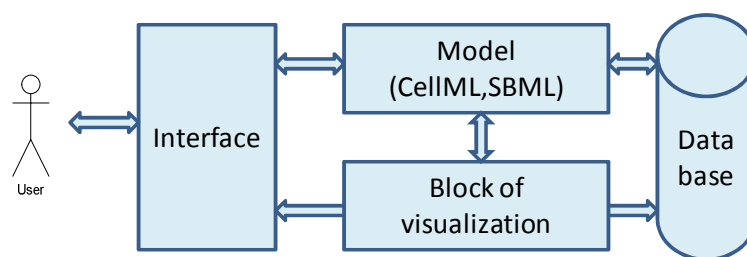


*Figure 6. The beginning of apoptosis process in the game*

*Figure 7. The destruction process of cell as a part of apoptosis process*

### **The system that converts the changes of model parameters into changes visual object model parameters**

The developing of a system for converting the changes of parameters in a model that is created by one of the modeling languages into changes in visual object model. The system may include an object generating system, which then react with each other in the scope of the model. This approach seems the most promising, but also the most difficult because of a lack of a comprehensive mathematical model of apoptosis as well as the complexity of the system. At present, the modeling environment can be created to display the model schematically. To create a full 2-D and 3-D graphical models that are close to photographic images and the ability to react to the changes of parameters will require significant effort. Fig 8 illustrates this approach.



*Figure 8. The interaction system of descriptive model and the rendering unit*

User – system user (player, student, doctor).

Interface – block that works with the user, implements receiving of model parameters and its visualization..

Database – system database that saves a set of images and its parameters, parameters scope, inconsistent values.

Block of visualization – rendering unit;

Model (CellML,SBML) – the model with a particular language that is created or selected by the user.

**Advantages** – high variability, the ability of process modeling and close connection to visualization.

**Disadvantages** – high complexity of implementation, absence of complex apoptosis models.

## Conclusion

Biological processes that occur in human body are complex to model and understand and not fully studied. A living organism is the most complex natural object accessible to our study. In this context, model visualization of the processes for the purpose and analysis is the natural way to improve the quality of their learning in training and in research.

Despite the relative abundance of methods and tools for modeling that were briefly discussed in this paper( SBMLEditor, libSBML, SBMLToolbox, MGSmodeller, OpenCell) today there are no full-featured environments providing not only modeling but also a graphical display of two-and three-dimensional models of biological processes. There are several approaches to create the visualization system of apoptosis models each of them have its own advantages and disadvantages:

The creation animated (video clips) or static images of processes by using descriptive (qualitative) model.

The creation a system for converting descriptive models with a limited set of parameters in a visual model (drawing, photography, picture, video clips) from a limited array of models.

The creation of system that converts changes in the model parameters that is created in one of the modeling languages (for example CellML) into changes in visual object model.

Literary sources indicate a huge interest of researchers to apoptosis modeling using the language CellML. Obviously using of CellML as a data source for the graphic model of the process is completely natural. Hence, further studies can offer implementation third approach with using OpenCell and CellML.

The work is funded by grant № 1288 from 2013 year for basic research in the natural sciences of Science Committee MES RK.

### References

- [1] Biological data visualization (2012) [http://en.wikipedia.org/wiki/Biological\\_data\\_visualization](http://en.wikipedia.org/wiki/Biological_data_visualization)
- [2] System Biology Markup Language (2012) - [http://sbml.org/Main\\_Page](http://sbml.org/Main_Page)
- [3] Modelling biological systems (2012) [http://en.wikipedia.org/wiki/Modelling\\_biological\\_systems](http://en.wikipedia.org/wiki/Modelling_biological_systems)
- [4] FieldML 0.2 Concepts and Serialisation. Richard Christie, Caton Little, Chris Bradley, Poul Nielsen, Randall Britten. Auckland Bioengineering Institute, The University of Auckland, New Zealand Updated 10 August 2010. [www.fieldml.org](http://www.fieldml.org).
- [5] FieldML. Physiome Project. <http://physiomeproject.org/software/fieldml/>.
- [6] Stefan Legewie, Nils Blüthgen, Hanspeter Herzel. Mathematical Modeling Identifies Inhibitors of Apoptosis as Mediators of Positive Feedback and Bistability. PLoS Computational Biology | [www.ploscompbiol.org](http://www.ploscompbiol.org) September 2006 | Volume 2 Issue 9.
- [7] CellML Model Repository (2013) - <http://models.cellml.org/cellml>
- [8] Bagci E.Z, Vodovotz Y., Billiar T.R., Ermentrout B., Bahar I. (2009) Computational insights on the competing effects of nitric oxide in regulating apoptosis. - <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2386238/>
- [9] Bagci E. Z., Vodovotz Y., Billiar T. R. et al (2006) Bistability in Apoptosis: Roles of Bax, Bcl-2 and Mitochondrial Permeability Transition Pores. Biophysical Journal, 90, 1546-1559. PubMed ID: 16339882.
- [10] Padmini Rangamani and Lawrence Sirovich (2007) Survival and apoptotic pathways initiated by TNF-alpha: modeling and predictions.
- [11] Computational Systems Neurobiology Group - Systems Biology of Neuronal Signalling (2013) <http://www.ebi.ac.uk/compneur-srv/SBMLeditor.html>.
- [12] libSBML (2013) - <http://sbml.org/Software/libSBML>
- [13] GNU Octave (2013) - [HTTP://WWW.GNU.ORG/SOFTWARE/OCTAVE/](http://www.gnu.org/software/octave/)
- [14] Keating, S. M., Bornstein, B. J., Finney, A., and Hucka M. (2006) SBMLToolbox: an SBML toolbox for MATLAB users. Bioinformatics, 22(10): 1275–1277.
- [15] MGSMoeller (2013) - [http://modelsgroup.bionet.nsc.ru/?page\\_id=491&lang=ru](http://modelsgroup.bionet.nsc.ru/?page_id=491&lang=ru)
- [16] Likhoshvai V. and Ratushny A. Generalized Hill function method for modeling molecular processes // Journal of Bioinformatics and Comp. Biology, 2007. V. 5. N.2. P. 521-531.
- [17] Казанцев Ф.В., Акбердин И.Р., Подколотный Н.Л., Лихошвай, В.А. Новые возможности системы MGSMoeller/ВаВилоВский журнал генетики и селекции, 2012, том 16, № 4/1.- [http://www.bionet.nsc.ru/vogis/pict\\_pdf/2012/16\\_4\\_1/10.pdf](http://www.bionet.nsc.ru/vogis/pict_pdf/2012/16_4_1/10.pdf)
- [18] OpenCell - <http://www.cellml.org/tools/opencell>
- [19] А.Т. Маншарипова, А. Али, З. А. Беркимбаева, Э.М. Хусаинова, Р.И. Мухамедиев, Е.Л. Мухамедиева // Под редакцией А.Т. Маншариповой

Апоптоз при сердечно-сосудистой патологии: молекулярно-генетические механизмы, регуляция и его визуализация. - Алматы, 2012. – 291 с.

**Authors:**

**R. I. Muhamedyev**, professor, head of CSSE&T of International IT University, Almaty, Kazakhstan ( Riga, Latvia).

**E. L.Muhamedyeva**, M.C.S., Riga (Latvia)

**A. T. Mansharipova**, professor, vice-rector of Kazakh-Russian Medical University, Almaty.

**Contacts:**

Ravil I. Muhamedyev, [email:ravil.muhamedyev@gmail.com](mailto:ravil.muhamedyev@gmail.com)

phones: +7 701 3393720,+7 777 2399788

## STORING BIG DATA USING NATURAL LANGUAGE ADDRESSING

Krassimir Markov, Krassimira Ivanova, Koen Vanhoof,  
Benoit Depaire, Vitalii Velychko, Juan Castellanos,  
Levon Aslanyan, Stefan Karastanev

**Abstract:** *Three characteristics define Big Data: volume, variety, and velocity [28]. An approach for storing Big Data is presented in the paper. It is called “Natural Language Addressing” [18; 19]. Its main idea is to use internal encoding of letters of a word or phrase as elements of co-ordinate vector which may be used as hyper-space address of the information connected to this word or phrase. As result the standard indexing and recompilation of information base are avoided and main problems of storing Big Data (Volume, Velocity, and Variety) can be solved.*

**Keywords:** *Big Data, Natural Language Addressing.*

### Introduction

Large unstructured or semi-structured datasets require a high level of computational sophistication because operations that are easy at a small scale — such as moving data between machines or in and out of storage, visualizing the data, or displaying results — can all require substantial algorithmic ingenuity. As a data set becomes increasingly massive, it may be infeasible to gather it in one place and analyze it as a whole. Thus, there may be a need for algorithms that operate in a distributed fashion, analyzing subsets of the data and aggregating those results to understand the complete set. One aspect of this is the challenge of data assimilation, in which we wish to use new data to update model parameters without reanalyzing the entire data set. This is essential when new waves of data continue to arrive, or subsets are analyzed in isolation of one another, and one aims to improve the model and inferences in an adaptive fashion — for example, with streaming algorithms [22].

The White House Office of Science and Technology Policy (OSTP) — in concert with several Federal departments and agencies — created the “Big Data Research and Development Initiative” to:

- Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data;
- Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen national security, and transform teaching and learning;
- Expand the workforce needed to develop and use Big Data technologies.

By improving the ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some the Nation's most pressing challenges [26].

For instance, as it is introduced in [13], the USA Department of Defense (DOD) is "placing a big bet on big data" investing \$250 million annually (with \$60 million available for new research projects) across the Military Departments in a series of programs that will:

- Harness and utilize massive data in new ways and bring together sensing, perception and decision support to make truly autonomous systems that can maneuver and make decisions on their own;

- Improve situational awareness to help war fighters and analysts and provide increased support to operations. The Department is seeking a 100-fold increase in the ability of analysts to extract information from texts in any language, and a similar increase in the number of objects, activities, and events that an analyst can observe.

Worldwide Big Data technology and services are expected to grow from EUR 2.4 billion in 2010 to EUR 12.7 billion in 2015 [15]. The challenge is to strengthen Europe's position as provider of innovative multilingual products and services based on digital content and data, addressing well identified industry and consumer market needs. Research and Innovation activities in this challenge will provide professionals and citizens with new tools to model, analyse, and visualise vast amounts of data from which to extract more value, to make an intelligent use of data coming from different sources and to create, access, exploit, and re-use all forms of digital content in any language and with any device [15].

The term "**Big Data**" applies to information that can't be processed or analyzed using traditional processes or tools. Increasingly, organizations today are facing more and more Big Data challenges. They have access to a wealth of information, but they don't know how to get value out of it because it is sitting in its most raw form or in a semi-structured or unstructured format; and as a result, they don't even know whether it's worth keeping (or even able to keep it for that matter).

Three characteristics define what IBM refers to as "Big Data": **volume**, **variety**, and **velocity** (Figure 1) [28].



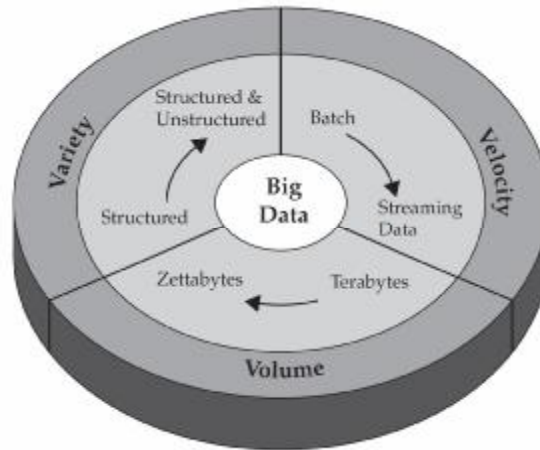


Figure 1. IBM characterizes Big Data by its volume, velocity, and variety or simply, V3 [28]

**Volume.** The sheer volume of data being stored today is exploding. In the year 2000, *800 000 petabytes* (PB) of data were stored in the world. Of course, a lot of the data that's being created today isn't analyzed at all and that's another problem we're trying to address with BigInsights. We expect this number to reach 35 zettabytes (ZB) by 2020. Twitter alone generates more than 7 terabytes (TB) of data every day, Facebook 10 TB, and some enterprises generate terabytes of data every hour of every day of the year. It's no longer unheard of for individual enterprises to have storage clusters holding petabytes of data [28].

We store everything: environmental data, financial data, medical data, surveillance data, and the list goes on and on. For example, taking your Smartphone out of your holster generates an event; when your commuter train's door opens for boarding, that's an event; check in for a plane, badge into work, buy a song on iTunes, change the TV channel, take an electronic toll route — everyone of these actions generates data.

**Variety.** The volume associated with the Big Data phenomena brings along new challenges for data centers trying to deal with it: its variety. With the explosion of sensors, and smart devices, as well as social collaboration technologies, data in an enterprise has become complex, because it includes not only traditional relational data, but also raw, semi-structured, and unstructured data from web pages, web log files (including click-stream data), search indexes, social media forums, e-mail, documents, sensor data from active and passive systems, and so on.

Quite simply, variety represents all types of data — a fundamental shift in analysis requirements from traditional structured data to include raw, semi-structured, and unstructured data as part of the decision-making and insight process. Traditional analytic platforms can't handle variety. However, an organization's success will rely on its ability to draw insights from the various kinds of data available to it, which includes both traditional and nontraditional [28].

**Velocity.** Just as the sheer volume and variety of data we collect and store has changed, so, too, has the velocity at which it is generated and needs to be handled. A conventional understanding of velocity typically considers how quickly the data is arriving and stored, and its associated rates of retrieval. While managing all of that quickly is good — and the volumes of data that we are looking at are a consequence of how quick the data arrives — we believe the idea of velocity is actually something far more compelling than these conventional definitions.

To accommodate velocity, a new way of thinking about a problem must start at the inception point of the data. Rather than confining the idea of velocity to the growth rates associated with your data repositories, we suggest you apply this definition to data in motion: The speed at which the data is flowing [28].

*The goal of this paper* is to present a new approach for storing Big Data. It is called “Natural Language Addressing” [18; 19]. Its main idea is to use internal encoding of letters of a word or phrase as elements of co-ordinate vector which may be used as hyper-space address of the information connected to this word or phrase. As result the standard indexing and recompilation of information base are avoided and main problems of storing Big Data caused by  $V^3$  (Volume, Velocity, and Variety) can be solved.

### **Resource Description Framework (RDF)**

The primary goal of Resource Description Framework (RDF) is to handle non regular or semi-structured data [1]. It is suitable for presenting and processing Big Data. RDF provides a general method to decompose any information into pieces called triples [20]:

- Each triple is of the form (“**Subject**”, “**Predicate**”, “**Object**”);
- Subject and Object are the names for two things in the world. Predicate is the relationship between them;
- Subject, Predicate, and Object, may be given as URI's (stand-ins for things in the real world);
- Object can additionally be raw text.

The power of RDF relies on the flexibility in representing arbitrary structure without a priori schemas. RDF offers the ability to specify concepts and link them together into a graph of data [9]. Each edge in the

graph is a single fact, a single statement, similar to the relationship between a single cell in a relational table and its row's primary key.

In this work we use datasets represented according the RDF rules. Datasets may be:

- *Small* RDF-datasets which contain up to several hundred thousand of RDF-instances;
- *Middle-size* RDF-datasets which contain from several hundred thousand up to 10 millions of RDF-instances;
- *Large* RDF-datasets may contain 50, 100, or more millions, as well as billions or trillions of RDF-instances.

### **Addressing**

The main idea of our approach is to use a special kind of addressing in multi-dimensional information spaces.

The common sense meaning of the concept “address” is such as a description of the location of a person or organization, as written or printed on mail as directions for delivery [35]; the conventional form by which the location of a building is described [8]; a sign in front of a house or business carrying the conventional form by which its location is described [30].

We will use the concept “*address*” in the sense accepted in the Computer Science: the code that identifies where a piece of information is stored [30]; a name or number used in information storage or retrieval that is assigned to a specific memory location; the memory location identified by this name or number; a name or a sequence of characters that designates an e mail account or a specific site on the Internet or other network [35].

It is important to take in account that the memory address may be of two kinds [34]:

- Physical location in any device (hard disk, main memory, flash memory);
- Logical (relative) location in a file given as an offset from the beginning of the file, i.e. the position of a byte in the file. In other words, it is the sequential number of the pointed byte in the file, starting from zero.

In this research we use concept “*memory address*” only in the second sense, i.e. as offset in a file stored somewhere in the computer accessible local or global network environment.

### **Name – Address - Route**

In January 1978, John F. Shoch, from “Xerox Palo Alto Research Center”, had written a very interesting note [16]. Later in the same year he had published this note in the paper [17]. This classical paper became as a mile stone in the further research concerning the naming,

addressing and routing at the first place with its “extremely general definition” [16]:

The “**name**” of a resource indicates “*what*” we seek,  
an “**address**” indicates “*where*” it is, and  
a “**route**” tell us “*how to get there*”.

This definition gives us a quick and intuitive understanding of the fundamental concepts of naming. Informally, a name is a string of symbols that identifies an object, thus both a human readable text-string and a binary number can be a name. Ideally, all objects would be named and handled in a uniform manner [24].

Shoch gave “some further detail to flesh this out” [16]:

I. A “name” is a symbol, usually a human-readable string, identifying some resource or set of resources. The name (what we seek) needs to be bound to the address (where it is).

II. An “address”, however, is the data structure whose format can be recognized by all elements in the domain, and which defines the fundamental addressable object. The address (where something is) needs to be bound to the route (how to get there).

III. A “route” is the specific information needed to forward a piece of information to its specified address.

*Thus, a “name” may be used to derive an “address”, which may then be used to derive a “route”.*

There is an interesting similarity between this structure and mechanisms used in programming languages (where one must bind a value to a variable), or in operating systems (where one must link a particular piece of code into a module) [16].

Establishing and supporting the correspondence between logical and physical addresses is duty of the operating systems or, in general, of all service functions of the local or global networks. This correspondence is transparent for the end user programs which request the name or logical address and operating environment is responsible to locate and access concrete physical location.

### **Natural Language Addressing**

In this research we follow a not published proposition of Kr. Markov to use the computer encoding of name (concept) letters as address of connected to it information stored in a multi-dimensional information (hyper) space. This way no indexes are needed and high speed direct access to the text elements is available.

This approach is called “Natural Language Addressing” (NL-Addressing).

Shoch’s definition [16] failed to capture that addresses are names too and names must eventually be mapped to routes [24]. In this sense,

the idea of NL-addressing is to use encoding of the name both as relative address and as route in a multi-dimensional information (hyper) space and this way to speed the access to stored information.

For instance, let have the next definition:

***“London: The capital city of England and the United Kingdom, and the largest city, urban zone and metropolitan area in the United Kingdom, and the European Union by most measures.”***

In the computer memory, for example, it may be stored in a file at relative address “00084920” and the index couple (**“London”**, **“00084920”**) means that at the memory address “00084920” the main text, *“The capital ... measures.”* will be stored.

To read/write the main text, firstly we need to find name “London” in the index and after that to access memory address “00084920” to read/write the definition.

If we assume that name “London” in the computer memory is encoded by six numbers (letter codes), for instance by using ASCII encoding system, than “London” is encoded as (76, 111, 110, 100, 111, 110). We may use these codes for direct access to memory, i.e. we have virtual index couple: (**“London”**, **“76, 111, 110, 100, 111, 110”**). In this case, we have written two times the same name. Because of this we may omit this couple and read/write directly from/to the address “76, 111, 110, 100, 111, 110”. For human this address will be shown as “London”, but for the computer it will be the co-ordinate array: (76, 111, 110, 100, 111, 110).

Now, what we need is a tool for storing and accessing information using Natural Language Addressing. At first glance, such tool may be the hash tables but the idea presented in this work differs from the hashing by two characteristics:

- The function which juxtapose the letters to integer numbers is one-one mapping and this way no collisions exist;
- This mapping (hash function) has to be used recursively for every symbol of a string to build hierarchical multi-layer set of hash tables and this way to speed the access to information.

For instance, the array “76, 111, 110, 100, 111, 110” may be assumed as a route to (or co-ordinates of) a point in a multi-dimensional (in this case: six-dimensional) information space and the definition may be stored in this point.

In other words, our function may be used recursively for every symbol and this way we will create hierarchical multi-layer set of hash tables. For the case of word “London” we will have six layers.

The natural language does not contain words only of six letters long. The length of the words is variable and in addition there exist names as phrases like “United Kingdom” or “European Union”. The set of

all natural words and phrases defines a multi-dimensional logical address hyper space with variable dimensions and unlimited size.

What we need are:

- A special algorithm which converts such multi-dimensional addresses in concrete routes to linear (relative) locations in the files (on the hard disk, for example);
- A program tool which will realize this algorithm.

### **Experimental program**

We have realized experimental program “**RDFArM**” for storing middle-size and large datasets of RDF triples and quadruples through Natural Language Addressing. Each RDF element can be stored by appropriate path, which is set by a natural language word or phrase.

No search indexes are needed and no recompilation of the data base is required after any update or adding new information in the data base.

Screenshots from the RDFArM program are shown at Figure 2 and Figure 3.

The main functions are **RDF-Write** and **RDF-Read** for which there are corresponded buttons. After storing the triples or quadruples, RDFArM displays two informative lines in the panel near to the “RDF-Write” button (Figure 2):

- Total time used for storing all given instances, in ticks (milliseconds);
- Average time used for storing of one instance.

After reading operation, in the window next to the RDF-Read button, two informative lines are shown (Figure 3):

- Total time used for extracting of all requested instances, in ticks (milliseconds);
- Average time used for extracting of one instance.

The time used is highly dependent on the possibilities of the operational environment and the speed of the computer hardware.

In the case of the Figure 2, 15472624 quadruple instances were stored for 63437758 milliseconds and one instance has been stored for average time of 4.1 milliseconds.

In the case of the Figure 3, 45595 quadruple instances were extracted for 151414 milliseconds and one instance has been extracted for average time of 3.3 milliseconds.

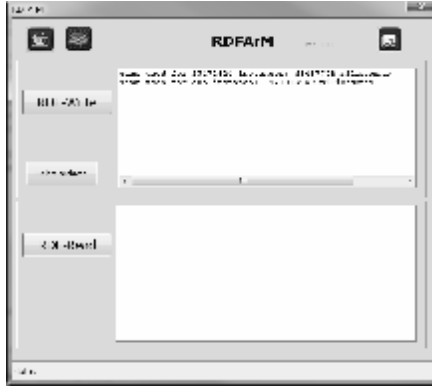


Figure 2. Content of RDFArM RDF-Write panel with informative lines



Figure 3. Content of RDFArM RDF-Read panel with informative lines

RDFArM is aimed to store RDF instances in dynamic perfect hash tables [21], i.e. the storing model is multi-layer. There are small differences between models for N-triples and N-quads.

N-triples can be stored in two variants. In the first, values of Predicates are names of the layers (archives), the Subjects are NL-addresses, and only Objects are saved. In the second, values of Subjects and Predicates are NL-addresses in the same archive, and only Objects are saved using couple (Subject, Relation) as NL-addresses (co-ordinates) of container where the Object is saved.

For N-quads, values of Subjects and Predicates are NL-addresses but Objects and Contexts are saved in different archives using couple (Subject, Relation) as NL-addresses (co-ordinates).

### Experiments

Our experimental environment includes program systems, computer configurations, datasets and experimental data like published benchmark results, different constants, ratio coefficients, etc. We have provided experiments with middle-size and large RDF-datasets, based on selected datasets from DBpedia's homepages and Berlin SPARQL Bench Mark (BSBM) to make comparison with known RDF triple stores. Experiments presented in this paper are based on:

- Real middle-size datasets from DBpedia's homepages: homepages-fixed.nt, geocoordinates-fixed.nt [10; 11];
- Artificial middle-size datasets from Berlin SPARQL Bench Mark (BSBM): BSBM 50K, BSBM 250K, BSBM 1M, BSBM 5M [7; 6];

– Real large datasets taken from DBpedia's homepages (infoboxes-fixed.nt) [DBpedia, 2007c], BSBM [6] and Billion Triple Challenge (BTC) 2012 [4];

– Artificial large datasets taken from Berlin SPARQL Bench Mark (BSBM): BSBM 25M, BSBM 100M [6].

The reason to make this choice is that we want to provide experiments with both real and artificial data. The artificial datasets like BSBM [3] contain standard artificially generated data and it is possible to adapt the software to have best results just for this kind of data. The DBpedia and BTC datasets were crawled using several seed sets collected from multiple real sources. Data in BTC datasets are encoded in N-Quads format.

The N-Quads is a format that extends N-triples with context. Each triple in N-Quad's document can have an optional context value [25]:

**<subject> <predicate> <object> <context>**,

as opposed to N-triples, where each triple has the form:

**<subject> <predicate> <object>**.

The notion of provenance is essential when integrating data from different sources. Therefore, modern RDF repositories store "subject-predicate-object-context" quadruples, where the context typically denotes the provenance of a given statement. The SPARQL query language can query such RDF datasets or entire collections of RDF graphs [33]. The context element is also sometimes used to track a dimension such as time or geographic location.

Applications of N-Quads include:

– Exchange of RDF datasets between RDF repositories, where the fourth element is the URI of the graph that contains each statement;

– Exchange of collections of RDF documents, where the fourth element is the HTTP URI from which the document was originally retrieved;

– Publishing of complex RDF knowledge bases, where the original provenance of each statement has to be kept intact.

N-Quads inherit the practical advantages of N-Triples:

– Simple parsing;

– Succinctness compared to alternatives such as reification or multi-document archives;

– Effective streaming and processing with line-based tools.

For example, a quadruple from BTC extracted from the data set [5] is given below:

<http://nektar.oszk.hu/resource/auth/magyar\_irodalom>  
<http://www.w3.org/2004/02/skos/core#narrower>  
<http://nektar.oszk.hu/resource/auth/hungarikum>  
<http://nektar.oszk.hu/data/auth/magyar\_irodalom>.



In this quadruple:

<subject> = <[http://nektar.oszk.hu/resource/auth/magyar\\_irodalom](http://nektar.oszk.hu/resource/auth/magyar_irodalom)>

<predicate> = <<http://www.w3.org/2004/02/skos/core#narrower>>

<object> = <<http://nektar.oszk.hu/resource/auth/hungarikum>>

<context> = <[http://nektar.oszk.hu/data/auth/magyar\\_irodalom](http://nektar.oszk.hu/data/auth/magyar_irodalom)>

Below we analyze comparison of RDFArM with three known systems – *Virtuoso*, *Jena*, and *Sesame*.

OpenLink **Virtuoso** is the first CROSS PLATFORM Universal Server to implement Web, File, and Database server functionality alongside Native XML Storage, and Universal Data Access Middleware, as a single server solution [27]. Virtuoso is a native triple store available in both open source and commercial licenses. It provides command line loaders, a connection API, and support for SPARQL and web server to perform SPARQL queries and uploading of data over HTTP. A number of evaluations have tested virtuoso and found it to be scalable to the region of 1B+ triples.

**Jena** is a Java toolkit for manipulating RDF models which has been developed by Hewlett-Packard Labs [2]. It has excellent support for RDQL queries, but does not provide an OKBC interface (however, given the level of RDQL support provided, the addition of an OKBC compatibility layer would be straightforward for the portion of the OKBC API that was implemented in the previous version of 3store). Jena is a java framework for building semantic web applications. Jena implements APIs for dealing with Semantic Web building blocks such as RDF and OWL. Jena's fundamental class for users is the Model, an API for dealing with a set of RDF triples. A Model can be created from the file system or from a remote file. Using JDBC, it can also be tied to an existing RDBMS such as MySQL or PostgreSQL.

**Sesame** is a de-facto standard framework for processing RDF data. This includes parsing, storing, inference and querying of/over such data. It offers an easy-to-use API that can be connected to all leading RDF storage solutions [32]. Sesame can be deployed on top of a variety of storage systems (relational databases, in-memory, file systems, keyword indexers, etc.), and offers a large scale of tools to developers to leverage the power of RDF and related standards. Sesame fully supports the SPARQL query language for expressive querying and offers transparent access to remote RDF repositories using the exact same API as for local access. Finally, Sesame supports all main stream RDF file formats, including RDF/XML, Turtle, N-Triples, TriG and TriX. Sesame is an open source framework for storage, inference and querying of RDF data. Sesame matches the features of Jena with the availability of a connection API, inference support, availability of a web server and

SPARQL endpoint. Like Jena SDB it provides support for multiple backends like MySQL and Postgres.

### Rank-based multiple comparison

Virtuoso, Jena and Sesame have several variants depending of database used. These variants have different loading times on the same computer configurations. In our comparisons we took the best result from the all benchmarks on given configuration.

We will use the Friedman test to detect statistically significant differences between the systems [14]. The Friedman test is a non-parametric test, based on the ranking of the systems on each dataset. It is equivalent of the repeated-measures ANOVA [31]. We will use Average Ranks ranking method, which is a simple ranking method, inspired by Friedman's statistic [23]. For each dataset the systems are ordered according to the time measures and are assigned ranks accordingly. The best system receives rank 1, the second – 2, etc. If two or more systems have equal value, they receive equal rank which is mean of the virtual positions that had to receive such number of systems if they were ordered consecutively each by other.

Let  $n$  is the number of observed datasets;  $k$  is the number of systems.

Let  $i_{ij}$  be the rank of system  $j$  on dataset  $i$ . The average rank for each system is calculated as

$$R_j = \frac{1}{n} \sum_{i=1}^k r_j^i.$$

The null-hypothesis states that if all the systems are equivalent than their ranks  $R_j$  should be equal. When null-hypothesis is rejected, we can proceed with the Nemenyi test [29] which is used when all systems are compared to each other. The performance of two systems is significantly different if the corresponding average ranks differ by at least the critical difference CD:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

where critical values  $q_\alpha$  are based on the Studentized range statistic divided by  $\sqrt{2}$ . Some of the values of  $q_\alpha$  are given in Table 1 [12].

Table 1. Critical values for the two-tailed Nemenyi test

systems	2	3	4	5	6	7	8	9	10
$q_{0.05}$	1.960	2.343	2.569	2.728	2.850	2.949	3.031	3.102	3.164
$q_{0.10}$	1.645	2.052	2.291	2.459	2.589	2.693	2.780	2.855	2.920

The results of the Nemenyi test are shown by means of critical difference diagrams.

Benchmark values for middle size datasets from our experiments and corresponded published experimental data from BSBM team are given in Table 2.

The ranks of the systems for the ten tests are presented below in Table 3.

Table 2. Benchmark values for middle size datasets

system	TEST									
	1	2	3	4	5a	5b	6	7	8	9
<b>RDFArM</b>	3	2272	14.79	3469	60	60	301	136412	1453	5901
<b>Sesame</b>	3	2404	19	2341	179	213	1988	21896	44225	282455
<b>Virtuoso</b>	2	1327	05	1235	23	25	609	7017	1035	3833
<b>Jena</b>	5	3557	13	3305	49	41	1053	70851	1013	5654

Table 3. Ranking of tested systems

system	ranks for the tests										average rank
	1	2	3	4	5a	5b	6	7	8	9	
<b>RDFArM</b>	2.5	2	3	4	3	3	1	4	3	3	<b>2.85</b>
<b>Sesame</b>	2.5	3	4	2	4	4	4	2	4	4	<b>3.35</b>
<b>Virtuoso</b>	1	1	1	1	1	1	2	1	2	1	<b>1.2</b>
<b>Jena</b>	4	4	2	3	2	2	3	3	1	2	<b>2.6</b>

All average ranks are different. The null-hypothesis is rejected and we can proceed with the Nemenyi test. Following [12], we may compute the critical difference by formula:

$$CD = q_{\alpha} \sqrt{\frac{k(k+1)}{6N}}$$

where  $q_{\alpha}$  we take as  $q_{0.10} = 2.291$  (from 0 [12; Table 5a]);  $k$  will be the number of systems compared, i.e.  $k=4$ ;  $N$  will be the number of datasets used in benchmarks, i.e.  $N=10$ . This way we have:

$$CD_{0.10} = 2.291 * \sqrt{\frac{4 * 5}{6 * 10}} = 2.291 * \sqrt{\frac{20}{60}} = 2.291 * 0.577 = 1.322$$

We will use for critical difference  $CD_{0.10}$  the value 1.322.

At the end, average ranks of the systems and distance to average rank of the first one are shown in Table 4.

Table 4. Average ranks of systems and distance to average rank of the first one

place	system	average rank	Distance between average rank of the system and average rank of the first one
<b>1</b>	<b>Virtuoso</b>	<b>1.20</b>	<b>0</b>
<b>2</b>	<b>Jena</b>	<b>2.60</b>	<b>1.4</b>
<b>3</b>	<b>RDFArM</b>	<b>2.85</b>	<b>1.65</b>
<b>4</b>	<b>Sesame</b>	<b>3.35</b>	<b>2.15</b>

The visualization of Nemenyi test results for tested systems is shown on Figure 4.

Analyzing these experiments we may conclude that RDFArM is at critical distances to Jena and Sesame. RDFArM is nearer to Jena than to Sesame. RDFArM, Jena, and Sesame are significantly different from Virtuoso.

Some recommendations to RDFArM may be given. RDF triple datasets has different characteristics depending of their origination. This causes the need to adapt RDFArM storage engine to specifics of concrete datasets. For instance, important parameters are length of strings and quantity of repeating values of subject, relation, and object.

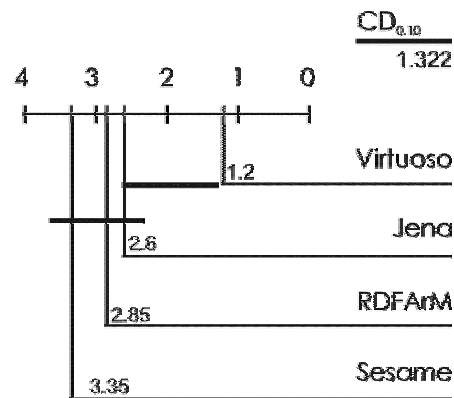


Figure 4. Visualisation of Nemenyi test results

### Graphical illustration

The graphic on Figure 5 is aimed to illustrate independence from size of the datasets. The storing time for BSBM 100M dataset is illustrated. RDFArM has constant time for storing of one triple independently of the number of already stored ones.

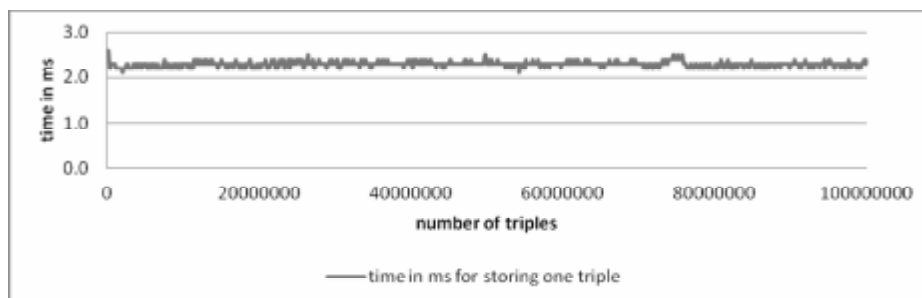


Figure 5. RDFArM storing time for one instance of BSBM 100M

## Conclusion

The term Big Data applies to information that can't be processed or analyzed using traditional processes or tools. Increasingly, organizations today are facing more and more "Big Data challenges". They have access to a wealth of information, but they don't know how to get value out of it because it is sitting in its most raw form or in a semi-structured or unstructured format.

Three characteristics define Big Data: volume, variety, and velocity. They have created the need for a new class of capabilities to augment the way things are done today to provide better line of site and controls over our existing knowledge domains and the ability to act on them [28].

An approach for storing Big Data was presented in the paper. It is called "Natural Language Addressing" [18; 19]. Its main idea is to use internal encoding of letters of a word or phrase as elements of co-ordinate vector which may be used as hyper-space address of the information connected to this word or phrase. As result the standard indexing and recompilation of information base are avoided.

This way, three main problems of storing Big Data may be solved:

- Volume – avoiding additional indexing, duplication of keywords, and corresponded pointers, leads to reducing additional memory needed for accessing information i.e. we may use addressing but not classical search engines;
- Velocity – avoiding recompilation of information base permits high speed of storing and immediately readiness of information to be accessed. This is very important possibility for stream data;
- Variety – natural language addressing permits creating a special kind of graph information bases which may operate both with structured as well as semi-structured information.

What gain and loss using NL-Addressing for storing Big Data?

The *loss* is additional memory for storing internal hash structures. But the same if no great losses we will have if we will build multiple balanced search trees or other kind of external indexing. It is difficult to compare with other systems because such information practically is not published.

The *benefit* is in two main achievements:

- High speed for storing and accessing the information;
- The possibility to update and access the information immediately after storing without recompilation the database and rebuilding the indexes. This is very important because many of the existing systems do not support updates.

## References

1. Andrae Muys, "Building an Enterprise Scale Database for RDF Data", Seminar, Netymon, 2007.
2. Apache Jena, [http://jena.apache.org/about\\_jena/about.html](http://jena.apache.org/about_jena/about.html) (accessed: 23.03.2013).
3. Berlin SPARQL Benchmark, <http://www4.wiwi.fu-berlin.de/bizer/BerlinSPARQLBenchmark/> (accessed 09.04.13).
4. Billion Triple Challenge 2012 Dataset <http://km.aifb.kit.edu/projects/btc-2012/> (accessed: 16.03.2013).
5. BTC data set from Datahub, <http://km.aifb.kit.edu/projects/btc-2012/datahub/data-0.nq.gz> (accessed: 16.03.2013).
6. Christian Bizer, Andreas Schultz, "The Berlin SPARQL Benchmark", In: International Journal on Semantic Web & Information Systems, Vol. 5, Issue 2, Pages 1-24, 2009, <http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/Bizer-Schultz-Berlin-SPARQL-Benchmark-IJSWIS.pdf>; see also <http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/> (accessed: 31.07.2013).
7. Christian Bizer, Andreas Schultz: Benchmarking the Performance of Storage Systems that expose SPARQL Endpoints; In: Proc. of the 4th International Workshop on Scalable Semantic Web knowledge Base Systems (SSWS2008), <http://www4.wiwi.fu-berlin.de/bizer/pub/BizerSchultz-BerlinSPARQLBenchmark.pdf> (accessed: 31.07.2013).
8. Collins English Dictionary – Complete and Unabridged, HarperCollins Publishers, 1991, 1994, 1998, 2000, 2003.
9. David C. Faye, Olivier Cure, Guillaume Blin. A survey of RDF storage approaches. Received, December 12, 2011, Accepted, February 7, 2012, ARIMA Journal, vol. 15 (2012), pp. 11-35.
10. DBpedia dataset "geocoordinates.nt" dated 2007-08-30, <http://wifo5-03.informatik.uni-mannheim.de/benchmarks-200801/geocoordinates-fixed.nt.gz> (accessed: 31.07.2013)
11. DBpedia dataset "homepages.nt" dated 2007-08-30, <http://wifo5-03.informatik.uni-mannheim.de/benchmarks-200801/homepages-fixed.nt.gz> (accessed: 31.07.2013)
12. Demsar J., "Statistical comparisons of classifiers over multiple data sets", J. Mach. Learn. Res., 7, 2006, pp. 1-30.
13. Fact Sheet: Big Data across the Federal Government, March 29, 2012, Office of Science and Technology Policy | Executive Office of the President, [http://www.whitehouse.gov/sites/default/files/microsites/ostp/big\\_data\\_fact\\_sheet\\_final.pdf](http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final.pdf) (accessed 09.04.13)
14. Friedman, M.: A comparison of alternative tests of significance for the problem of m rankings, Annals of Mathematical Statistics, Vol. 11, 1940, pp.86-92.
15. HORIZON 2020 – WORK PROGRAMME 2014-2015. [http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014\\_2015/main/h2020-wp1415-leit-ict\\_en.pdf](http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/main/h2020-wp1415-leit-ict_en.pdf) (accessed: 29.12.2013)

16. John F. Shoch "A note on Inter-Network Naming, Addressing and Routing", Xerox, Palo Alto, Research Center, Palo Alto - California 94305, USA, January 1978. <http://www.postel.org/ien/pdf/ien019.pdf> (accessed: 21.02.2013)
17. John F. Shoch. Inter-Network Naming, Addressing, and Routing. In Proc. of the Seventeenth IEEE Conference on Computer Communication Networks, pages 72–79, Washington, D.C., 1978.
18. Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, "Introduction to the Natural Language Addressing", International Journal "Information Technologies & Knowledge" Vol.7, Number 2, 2013, ISSN 1313-0455 (printed), 1313-048X (online), pp. 139–146.
19. Krassimira Ivanova, Vitalii Velychko, Krassimir Markov, "Storing RDF Graphs using NL-addressing", In: G. Setlak, M. Alexandrov, K. Markov (ed.), Artificial Intelligence Methods and Techniques for Business and Engineering Applications. ITHEA® 2012, Rzeszow, Poland; Sofia, Bulgaria, ISBN: 978-954-16-0057-3 (printed), ISBN: 978-954-16-0058-0 (online), pp. 84 – 98.
20. Mario Briggs "DB2 NoSQL Graph Store", What, Why & Overview, A presentation, Information Management software IBM, 2012, [https://www.ibm.com/developerworks/mydeveloperworks/blogs/nlp/resource/DB2\\_NoSQLGraphStore.pdf?lang=en](https://www.ibm.com/developerworks/mydeveloperworks/blogs/nlp/resource/DB2_NoSQLGraphStore.pdf?lang=en) (accessed: 01.12.2012)
21. Martin Dietzfelbinger, Anna Karlin, Kurt Mehlhorn, Friedhelm Meyer auf der Heide, Hans Rohnert, and Robert E. Tarjan, 1994. Dynamic Perfect Hashing: Upper and Lower Bounds. SIAM J. Comput, 23, 4 (Aug. 1994), ISSN: 0097-5397. pp. 738-761. <http://portal.acm.org/citation.cfm?id=182370#> (accessed: 20.07.2013).
22. National Research Council, "The Mathematical Sciences in 2025", Washington, DC: The National Academies Press, USA, 2013. ISBN-13: 978-0-309-28457-8. [http://www.nap.edu/catalog.php?record\\_id=15269](http://www.nap.edu/catalog.php?record_id=15269) (accessed 09.04.13).
23. Neave, H., Worthington, P.: Distribution Free Tests. Routledge, 1992.
24. Nick Jording and Flemming Andreassen "A Distributed Wide Area Name Service for an Object Oriented Programming System", DIKU, Department of Computer Science, University of Copenhagen, Denmark, 1994.
25. N-Quads: Extending N-Triples with Context <http://sw.deri.org/2008/07/n-quads/> (accessed: 16.03.2013).
26. Obama Administration unveils "BIG DATA" INITIATIVE: Announces \$200 Million in New R&D Investments, Office of Science and Technology Policy | Executive Office of the President. March 29, 2012. [http://www.whitehouse.gov/sites/default/files/microsites/ostp/big\\_data\\_press\\_release\\_final\\_2.pdf](http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf) (accessed 09.04.13)
27. OpenLink Virtuoso Universal Server: Documentation <http://docs.openlinksw.com/pdf/virtdocs.pdf>, <http://virtuoso.openlinksw.com/> (accessed: 23.03.2013).
28. Paul C. Zikopoulos, Chris Eaton, Dirk de Roos, Thomas Deutsch, George Lapis "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data", Copyright© 2012 by The McGraw-Hill Companies, ISBN 978-0-07-179053-6, MHID 0-07-179053-5, 2012, 166 p.

29. Peter Nemenyi, "Distribution-free multiple comparisons Unpublished", PhD thesis; Princeton University Princeton, NJ, 1963.
30. Princeton University "About WordNet", WordNet, Princeton University, 2010 <http://WordNet.princeton.edu> (accessed: 23.07.2012).
31. R. A. Fisher, Statistical methods and scientific inference (3rd edition), Hafner Press, New York, 1973, ISBN 978-002-844740-7.
32. Sesame, OpenRDF, <http://www.openrdf.org/index.jsp> <http://www.openrdf.org/doc/sesame2/2.3.2/users/userguide.html#chapter-sesame2-whats-new> (accessed: 01.12.2012).
33. SPARQL Query Language for RDF "W3C Recommendation", 2008, <http://www.w3.org/TR/rdf-sparql-query/> (accessed: 23.03.2013).
34. Stably D., "Logical Programming with System", 360, New York, 1970
35. The American Heritage® "Dictionary of the English Language" Fourth Edition copyright© 2000 by Houghton Mifflin Company, Updated in 2009; Published by Houghton Mifflin Company. All rights reserved.

#### **Authors:**

**Krassimir Markov** – Assoc. Prof. Dr.; Institute of Mathematics and Informatics, BAS, Bulgaria; e-mail: [markov@foibg.com](mailto:markov@foibg.com);

**Krassimira Ivanova** – Assist. prof. Dr.; University of National and World Economy, Sofia, Bulgaria; Institute of Mathematics and Informatics, BAS, Bulgaria; e-mail: [krasy78@mail.bg](mailto:krasy78@mail.bg);

**Koen Vanhoof** – Prof. Dr., Hasselt University; Campus Diepenbeek; Department of Applied Economic Sciences, Agoralaan; BE-3590 Diepenbeek; Belgium; e-mail: [koen.vanhoof@uhasselt.be](mailto:koen.vanhoof@uhasselt.be);

**Benoit Depaire** – Assoc. prof. Dr.; Hasselt University; Campus Diepenbeek; Department of Applied Economic Sciences, Agoralaan; BE-3590 Diepenbeek, Belgium; e-mail: [benoit.depaire@uhasselt.be](mailto:benoit.depaire@uhasselt.be)

**Vitalii Velychko** – Assoc. prof. Dr.; Institute of Cybernetics, NASU, Kiev, Ukraine; e-mail: [aduis@rambler.ru](mailto:aduis@rambler.ru);

**Juan Castellanos** – Assoc. prof. Dr.; Head of Natural Computing Group, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s.n., 28660 Boadilla del Monte, Madrid, Spain; e-mail: [jcastellanos@fi.upm.es](mailto:jcastellanos@fi.upm.es);

**Levon Aslanyan** – Prof. Dr.; Institute for informatics and automation problems of NAS RA, 1, P. Sevak street, Yerevan 0014, Armenia, e-mail: [lasl@sci.am](mailto:lasl@sci.am);

**Stefan Karastanev** – Assist. prof.; Institute of Mechanics, BAS, Bulgaria; e-mail: [stefan@imbm.bas.bg](mailto:stefan@imbm.bas.bg)



# A THEORY AND APPROACH FOR DEVELOPING CLOUD APPLICATIONS

Danko Naydenov

**Abstract:** *In this paper it will be consider a methodology that can be used to create cloud applications. There are many possible approaches, but this was chosen because it is ease to be used and ease to be accessed. System requirements overview is made. For selected approach it is considered some application programing interfaces for user maintenance and authentication, storing data and deploying the application. It is discussed how standard technologies can be apply to this service also an overview of some security requirements is made.*

**Keywords:** *Cloud computing, PaaS, Cloud application*

## Introduction

With the development of information technology users and developers turn to cloud services. The reasons for this are several, namely the low cost of initial investment, high availability and reliability of applications, the low cost of maintenance and others[6].

## Goals and objectives

The goal that sets this paper is to demonstrate a methodology for developing cloud applications.

To achieve this goal, it is defined a task to develop a sample application representing a chat room. This application has to be designed for cloud environment. It has to be deployed there and works in such an environment.

## Analysis of the problem

On the market there are many cloud vendors and various cloud services which can be grouped into three main models[6]:

Infrastructure as a Service (IaaS)

Platform as a Service (PaaS)

Software as a Service (SaaS)

The most appropriate service for the development of applications is PaaS. The reason for this is that the approach which is been used to create a cloud application largely resembles the development of a standard WEB application.

For provider is selected Google and offering by them PaaS - Google App Engine. The choice is based on the following factors:

- because cloud computing is a relatively new technology, and to offer such a service requires a significant investment, not a variety of providers are available and participants are giants like Microsoft, Amazon, Google and others;

- Google prevailed in the election because they currently offer free access to develop, test and deploy cloud applications;
- cloud service provided by Google offers the possibility to use several languages, but this paper will focus on Java[3];

### **System requirements and installation of work environment**

It is assumed that the developer has installed and working version of Java[3]. Because the Google App Engine works using Java 7, cloud service provider recommends using the same version to build the application.

To develop cloud application Google offer App Engine Java software development kit (SDK). This is an environment that emulates all aspects of the cloud, which however, works locally and allows the used to develop and test the application.

App Engine Java SDK is a standalone application that does not have a high degree of automation. Eclipse[2] is one of the most popular development environments for Java[3] applications, and because of that Google offers Plugin for it. Developing application for Google App Engine is easy and similar to the development of servlet[5] application. By using the Google Plugin this task is even easier because it allows all actions to be performed directly from the development environment.

### **Installing the Google Plugin for Eclipse**

At the time of development of this paper the current version of the development environment is Eclipse Kepler (4.3.2). To install the Plugin user have to perform the following actions:

1. Choose from „Help” menu „Install New Software...”
2. In newly open window in field „Work with” is entered:

<https://dl.google.com/eclipse/plugin/4.3>

Next activity is pressing the button „Add...” right to the field. Confirm with „OK” for the new window. Field „Name” leave blank, as it will be filled later by page of Plugin.

3. After loading the table, press the icon to expand two headers next to "Google Plugin for Eclipse" and "SDKs". Place a check mark in the boxes with the following elements: "Google Plugin for Eclipse 4.3" and "Google App Engine Java SDK". The user has to ensure that a checkbox next to "Contact all update sites during install to find required software" is selected. Press the "Next>" and follow the instructions on the installation.

4. When installation is complete, the working environment Eclipse will prompted to reboot. After restarting the environment new additions will be active and available.

### **Creating a Java App Engine project**

Google have chosen as one of its cloud platforms service to be Java Servlet[5] technology. This determines the manner in which the application will interact with the server.

To create a new project from "File" menu select "New / Other ...". From the opened dialog choose the section "Google" and then "Web Application Project". After pressing the "Next>" a dialog is opened to define a new project. As name of the project can be set "ChatRoom" and for the package name "com.cc.chatroom". Additional settings will be: deselect a "Use Google Web Toolkit" and select "Use Google App Engine". After clicking the "Finish" button the structure of the sample project is automatically created. Main content is divided in two directories:

- src – contains application code files;

- war – contains the compiled application in WAR format (WAR archive is not supported)

The newly created application can be launched from the menu "Run" by selecting "Debug As> Web Application", and it can be accessed via a standard browser at:

<http://localhost:8888/>

### **Authentication of users**

By creating a standard project it is also created a servlet with the name of the project, which is registered in the "web.xml" file. In this example it is in "ChatRoomServlet.java". This is a class that inherits "javax.servlet.http.HttpServlet" class. Initially, the effects of this class is to handle a GET request and display as text message "Hello, world".

This servlet will be modified so that a POST request is forward to be processed by function processing GET request.

Authenticating the users is a major task for any application and can be very complex. When creating applications for Google App Engine it is possible to use the authentication provided by Google. This means that instead writing a code in application that is being developed to take care of the maintenance of users, this task can be left to Google. For this purpose, if a user wants to work with the application it must be registered and have a valid Google account. Code that makes authentication in this case is rather short:

```
public class ChatRoomServlet extends HttpServlet {
    public void doPost(HttpServletRequest req, HttpServletResponse resp)
        throws IOException {
        doGet(req, resp);
    }
    public void doGet(HttpServletRequest req, HttpServletResponse resp)
        throws IOException {
        UserService userService = UserServiceFactory.getUserService();
```

```

User user = userService.getCurrentUser();
if (user != null) {
    resp.getWriter().println(
        "<p>Welcome " + user.getNickname()
        + ". To sign out click " + "<a href="
        + userService.createLogoutURL(req.getRequestURI())
        + ">here</a>.</p>");
} else {
    resp.getWriter().println(
        "<p>Welcome to ChatRoom. You can sign in from <a href="
        + userService.createLoginURL(req.getRequestURI())
        + ">here</a></p>");
}
}
}

```

UserService is interface that gives access to the main tasks of identifying the user. `getCurrentUser()` returns an object for current user. If this object is null, this means that there is no currently registered user. With `userService.createLoginURL( req.getRequestURI())` browser is redirected to Google page for user authentication. When using a local server for development and testing, it is possible cases when Internet is not available and in the cases the browser is redirected instead of Google login form to a standard form where the authentication is always possible only with e-mail address. The parameter that has been passed to the function `createLoginURL`, is an address which will be used to redirect the browser after successful authentication. For logout it is used other function `createLogoutURL`, and the parameter again is return address.

The logic of demonstrating code is following:

- if there is currently authenticated user it is displayed the message "Welcome" followed by the name of the user. Then an unsubscribe link is present;
- if there is no currently authenticated user it is displayed "Welcome to ChatRoom", following by the link to authenticate the user;

### Dynamic content

Using servlets allows generation of dynamic content in pages, but unfortunately when there is need to generate more complex pages, the task can become quite difficult. To solve this problem it is developed multiple template systems for Java. These are systems that combine static HTML content and dynamically changing fragments. Thus complex layouts on a page can be generated separately and independently, even using assistive environments and tools, and specific content can be added at the designated places. In this paper we will look at JavaServer

Pages (JSPs)[1], as one of the most popular systems, at the same time supported by Google App Engine.

By default, all files with the extension .jsp from the war directory and all subdirectories except WEB-INF are mapped to the corresponding URLs.

In order to use JSPs, Eclipse must be configured to use the JDK, not the JRE. This can be done from the menu: Window > Preferences > Java > Installed JREs.

Similar result to that obtained from already discussed servlet can be obtained from JSP template. This is a file that can be placed in the war directory and and be given the name index.jsp. Although the file is named index.jsp this does not mean that it will be loaded by default. To become this true a modification is required for the configuration file web.xml located in the war / WEB-INF directory. There is need to update the contents of the tag <welcome-file>:

```
<welcome-file-list>
  <welcome-file>index.jsp</welcome-file>
</welcome-file-list>
```

The main fragment of code that performs authentication of users using JSP templates is as follows:

```
...
<body>
  <%
    UserService userService = UserServiceFactory.getUserService();
    User user = userService.getCurrentUser();
    if (user != null) {
      %>
      <p>Welcome <%=escapeHtml3(user.getNickname()) %>. To sign out click <a href="
      <%=userService.createLogoutURL(request.getRequestURI())%>">here</a>.</p>
      <%
    } else {
      %>
      <p>Welcome to ChatRoom. You can sign in from <a href="
      <%=userService.createLoginURL(request.getRequestURI())%>">here</a>.</p>
      <%
    }
  %>
</body>
...
```

An important thing that needs to be addressed is the output of the text on the screen of the browser. In this case it is the user's name. If it works as demonstrated in servlet user.getNickname(), this can lead to security problems. The reason for this is that if a malicious user as its name defines some HTML code, it will be put in the output, and will be interpreted by the browser. To avoid security problems of this nature it is recommended that all strings which will be displayed to the screen and the application does not have control of there content, to be escaped.

There is a standard library Apache - Commons Lang[7], which has a similar function. To use this feature in application there is need to place the library file in war\WEB-INF\lib directory and include the library that contains the function with following line:

```
<%@page import=" org.apache.commons.lang3.StringEscapeUtils. escapeHtml3"%>
```

The function is named `escapeHtml3`, and by using it all strings is converted so that they are displayed as text in the browser.

As a result, it is possible to display any text to the browser, even one that contains some HTML, and it will be transformed into the escape sequence:

```
<%=escapeHtml3(user.getNickname()) %>
```

Another important part of the `web.xml` file is association of servlet to a physical address. If we define a new servlet in the application this does not automatically make it visible and accessible to the client browser. In order to make this servlet accessible via some address and to receive GET or POST requests this servlet need to be specified in the configuration file `web.xml`. Fowling example will demonstrate how this can be done for servlet that will receive messages sent to the server by users. Servlet will be defined in the file `SendMessageServlet.java`.

```
<servlet>
  <servlet-name>send</servlet-name>
  <servlet-class>com.cc.chatroom.SendMessageServlet</servlet-class>
</servlet>
<servlet-mapping>
  <servlet-name>send</servlet-name>
  <url-pattern>/send</url-pattern>
</servlet-mapping>
```

First section `<servlet>` describes the name and which class will process requests.

Second section `<servlet-mapping>` specifies the address where the servlet will be available.

So define and describe servlet will be activated from a HTML form which will be placed in `index.jsp`:

```
<form action="/send" method="post" id=frm>
  <pre id="output" style=
    "height: 300px; width: 600px; overflow: auto; border: 1px solid black;">
  </pre>
  <input id="input" name="input" type="text" maxlength=300
    style="width: 600px"/><br/>
  <input type="submit" value="Send"/>
  <script type="text/javascript">
    document.forms.frm.input.focus();
  </script>
</form>
```

The method that have been used to send request to servlet is POST and `action="/send"` set the address where the servlet is available.

### **Data handling**

Cloud infrastructure is quite different architecture from normal PCs and even servers. If an application for some time communicates with one node from the cloud infrastructure, then the next requests can be redirected to another node. Therefore, it is difficult to determine exactly where to save the data so that they are available for next call. In addition there are actions for backup the data in case of failure so that they can be restored in such cases.

To ignore all the details related to the storage and maintenance of data, Google App Engine offers a service called Datastore. This is a set of functions enabling storing and retrieving the data using simple API.

In the Datastore recorded unit of information is called entity. Each entity consists of a key and value. The key must be unique and is used to identify the entity. The value can be any object.

When the entity has been saved in Datastore this lead to number of actions related to the sharing of data between data centers, backup, and many other activities related to cloud architecture, all of which is hidden from the developer.

To save a message which had been sent to chat room, it is necessary to modify the servlet SendMessageServlet like this:

```
String content = req.getParameter("input");
Date date = new Date();
Entity message = new Entity("Message");
message.setProperty("user", user);
message.setProperty("timestamp", date);
message.setProperty("text", content);
DatastoreService datastore = DatastoreServiceFactory.getDatastoreService();
datastore.put(message);
resp.sendRedirect("/index.jsp");
```

- the first row retrieve entered text by the user;
- second line creates an object representing the current system time which is used as a characteristic of the message;
- third line creates an item that will be recorded in the Datastore. This element is of type "Message", and don't contain a key value. After saving the item a unique key value will be assigned;
- each element can have any number of properties. In describe case these three properties are: user, timestamp of message and the message itself (lines 4 to 6);
- actual recording is on row number eight: `datastore.put(message);`
- and final task is to redirect the browser to the default template;

Retrieving of data is related to the execution of the following sequence:

- creating an object which define the search query;
- create a list of items corresponding to the request;

In the code that looks like this:

```
DatastoreService datastore = DatastoreServiceFactory
    .getDatastoreService();
Query query = new Query("Message").addSort("timestamp",
    Query.SortDirection.ASCENDING);
Iterable<Entity> messages = datastore.prepare(query).asIterable();
```

Object query retrieves all elements of type "Message", using the characteristic "timestamp" to perform sorting.

messages is an object containing all the elements meeting the request.

Access to individual fields can be done as follows:

```
for (Entity message : messages) {
    ...
    ...escapeHtml3(message.getProperty("text").toString())...
    ...
}
```

Again, it should be kept in mind security, by simply using the function to convert special characters, if any, in the entered text.

Sending a message is made by submitting the page and passing the control to the servlet processing this request. Receiving information from the server side is more complicated. If this is done again with submit this can lead to loss of Writer current message. It is therefore necessary to use an asynchronous request to the server - Ajax[4]. In this case, the background will ask a server for current messages from chat room. After receiving the response all messages will be recorded in the appropriate controls. This whole concept is implemented using JavaScript as follows:

```
var req = null;
if (window.XMLHttpRequest) {
    req = new XMLHttpRequest();
} else if (window.ActiveXObject) {
    req = new ActiveXObject("Microsoft.XMLHTTP");
}
function processRequest() {
    if (req.readyState == 4 && req.status == 200) {
        var message = req.responseText;
        if (message != null && message != "") {
            var outputElement = document.getElementById("output");
            outputElement.innerHTML = message;
            outputElement.scrollTop = outputElement.scrollHeight;
        }
    }
}
function doReceive() {
    req.onreadystatechange = processRequest;
    req.open("POST", "getMessage", true);
    req.setRequestHeader("Content-type", "application/x-www-form-urlencoded");
    req.send();
}
```



req variable is an object serving to organize asynchronous communication. doReceive is a function that specifies which function will handle the response to the query, specifies the servlet which will handle the queries and what is the type of request, and makes the actual call. When a response is receipt the function processRequest is activated to handle this particular answer. After checks for correctness of response and data the response is written as text in graphic control which contains all posted messages.

### **Deploying the application**

The last act of creating a cloud application is to be deployed in cloud infrastructure so it can be available over the Internet. To do so an application ID has to be created. This identifier can be setup from:

<https://appengine.google.com/>

To do this you have to own a valid account for Google. After loading the portal for administrating the applications the button "Create an Application" has to be pressed. The registered identifier is using to access the application from the Internet. If a free version of the domain is used and assume that the identifier is chatroom, then the application can be accesses from:

<http://chatroom.appspot.com/>

Deploying the finished application can be done from Eclipse environment. For this purpose it is necessary to edit the file appengine-web.xml and set the element <application> with registered identifier, for this case chatroom. Actual deployment can be done from the toolbar by pressing the Google button and then the option "Deploy to App Engine...". There are specifying issues for Google account and some options about loaded application into cloud infrastructure.

### **Conclusion**

The process of creating and deploying a cloud Application for Google App Engine largely resembles the same process for standard Java WEB application, although there are some small differences. The main difference is that cloud applications do not have a standard relational database. Cloud provides a technology for storing data that can be defined as an objective database. This despite the initial stress in developers is pretty powerful mechanism for storing data that requires some readjustment in thinking and structuring of an application.

### **References**

[1] Hans Bergsten; Java Server Pages (December 2000); Publisher: O'Reilly Media

- [2] Frank Budinsky, David Steinberg, Ed Merks, Raymond Ellersick; Eclipse Modeling Framework: A Developer's Guide (August 2003); Publisher: Addison-Wesley Professional
- [3] David Flanagan; Java in a Nutshell (March 2005); Publisher: O'Reilly Media
- [4] Jesse James Garrett; Ajax: A New Approach to Web Applications (February 2005)
- [5] Jason Hunter; Java Servlet Programming (November 1998); Publisher: O'Reilly Media
- [6] Danko Naydenov, Modern paradigm and trend in Cloud Computing (Jun 2012); Informatics in the scientific knowledge 2012
- [7] <http://commons.apache.org/lang/>

**Author:**

**Danko Naydenov Naydenov**, TU-Varna  
0899 066 331, [naydenov@tu-varna.bg](mailto:naydenov@tu-varna.bg)

## GEOPHYSICAL RESEARCH DATA PREPROCESSING MODULE

P. Gritsenko, Y. Kuchin, R. Muhamedyev,  
Z. Nurushev, K. Yakunin

**Abstract:** *A software module was developed for preprocessing of electric logging data of wells at uranium deposits, at which underground leaching extraction method is used. The software implements a number of preprocessing algorithms. Among implemented functions are: processing functions for arbitrary data sets, specific functions for a given set of data and tools used in the preliminary analysis of the data. Application of this package can improve recognition quality by 5-10%.*

**Keywords:** *preprocessing, learning samples, uranium deposit, recognition process, machine learning*

### Introduction

To assess the suitability of the well and for justified execution of technological processes of uranium extraction, geophysical research is performed, which usually evaluates the electromagnetic properties of rocks.

Data obtained in the course of the geophysical research can be interpreted by a specialist or an intelligent system based on machine learning methods [1]. Note that systems of automatic recognition and classification are used in many spheres of human activity, including processes of solving problems of lithology. [2, 3, 4]. Machine-learning techniques have been studied in detail. In some cases it's possible to achieve results, which are close to optimal, on a given set of data. However, in the case of using the methods, it is essential to select classification features and preprocess data. In case of geophysical logging data, classification features are initially defined. These features are IK, KS, PS and depth. However, the quality of the results depends strongly on preprocessing methods [5, 6].

For solving the data preprocessing problem for electric well logging of uranium deposits with underground leaching extraction method, "Preprocessing Module" software was developed implementing a number of preprocessing algorithms, and tools for data analysis and metric recognition algorithm (K Nearest Neighbors method).

"Preprocessing Module" includes:

- 1) The functions of processing arbitrary data sets (norm(), clean\_anomalies(), clean\_noisy\_points(), window()), which are not tied to specific features of electric logging data.

2) Processing functions that are specific to the electric logging data of uranium wells (`shift_corr()`, `inverse_PS()`, `align_level()`). The practical usefulness of these functions when used on other data is questionable.

3) Tools for preliminary data analysis (`crude_define()`, `for_matlab()`, `count_anomalies()`, `error_percent_per_soil()`, `divide_to_classes()` and others) , which present opportunities for research, analysis, output of the data in various formats for use in other program.

### **Main problems with geophysical research data**

Although machine-learning techniques are well researched, developed and implemented in several software libraries/applications, such as Weka, RapidMiner, Orange, etc., neither of them are able to solve arbitrary practical recognition problem out of the box. The main reason is, that in every practical recognition problem different problems with data can be showed up, which wouldn't allow Machine Learning algorithms to reach desired quality of recognition. Some of the problems are common for many different types of data (such as non-normalized data, noise and faults of measurement), other are specific for the given practical problem.

Main problems with the geophysical research data, detected during the research are:

- Noise in signals – equipment for measurement is imperfect, which leads to substantial amount of noise in the signal. Also, interference from drilling equipment is also being recorded during the measurements.
- Anomalies in signals – technological process of well logging at uranium deposits of Kazakhstan has several peculiarities, which can lead to anomalies in data. For example, in certain cases IK (inductivity) parameter can be measured for metal constructions of wells, rather than for soil, which would lead to abnormal value of the measured parameter and serious problems during recognition.
- Different recording level of signals – geophysical research of different wells of one uranium deposits can be performed using different equipment, or same equipment with slightly different configuration, which leads to different recording level of signals. Recording level of electrical logging depends on many factors: the diameter of the tool, probe length, diameter of the well, mineralization of solute, correct calibration, etc. For one group of well range of one parameter can be in range of [10; 30] units, for another it can be [20; 50], while geological structure is similar. This has negative impact on recognition, as well as on normalization of the data.
- PS inverse – PS is one of the parameters measured during geophysical research at uranium deposits of Kazakhstan. Basically, it

measures natural electromagnetic field properties of the soil. Recording level and polarity depends on the ratio of mineralization of drilling solute and groundwater; increased salinity of groundwater in some horizons leads to a change in polarity of PS signal. The problem has to be addressed by bringing the parameter to one polarization for all training set.

- Non-normalized data – different parameters measured during geophysical research can have significantly different value range, which is a problem for most recognition methods – most methods, such as artificial neural networks are well suited for data, normalized in [-1; 1] range.

- Noisy samples – training sample used during the research was obtained during interpretation of the geophysical research data by several experts-geologists. Human factors, as well as contradictions in opinions among experts leads to noise in the sample – i.e., objects in certain compact areas of feature space are recognized differently by different experts, which is a problem for automated recognition using machine learning algorithms.

- Shift of signals – measurement of different parameters of one well is usually performed two or more separate tools. Due to inaccuracy, the signals obtained for separate tools can be shifted relating to one another. However, as the all the measurements used during the research are of electromagnetic nature, they are correlated to some extent, which means that it is possible to fix the shift between several signals.

All of the problems were addressed in the developed Preprocessing Module – it contains functions for both training sample and data to be recognized, which successfully solve the problems.

### **Preprocessing algorithms developed in the module**

Table 1 shows problems and methods implemented in the Preprocessing Module, that solve the problems.

*Table 1. Problems of data and methods for solving them*

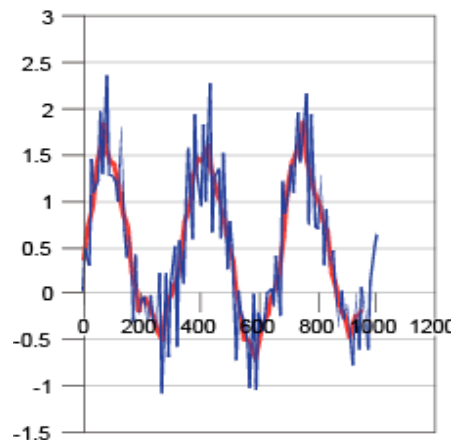
Data Preprocessing	
Noise in signals	Wavelet analysis
Anomalies in signals	Fixing anomalies
Different recording level of signals	Bringing levels to one level
PS inverse	Data recovery
Non-normalized data	Data normalization
Noisy samples	Removing noise points
Shift of logs	Removing log shift

**Wavelet analysis** allows to:

- Localize special points;
- Analyze frequency and signal amplitude variability;
- Identify anomalies in function and its derivatives;
- Calculate fractal characteristics of signals.

In addition to data normalization, in order, to improve quality of recognition various methods of smoothing, allowing eliminating noise and other components of uninformative data, can be used. Usually, there are two kinds of discrete transforms used as smoothing methods: Fourier analysis and Wavelet transform.

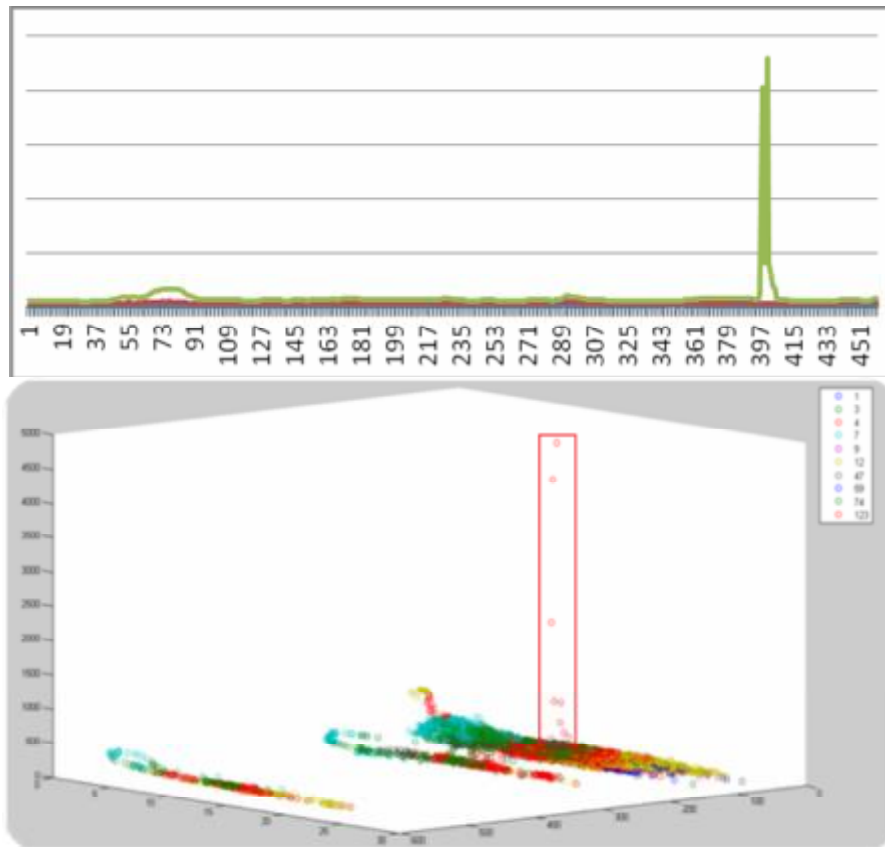
Since data recording occurs at a specified interval, there is low possibility to face high frequency noise in log data. Moreover, elimination of high-frequency component will lead to information loss (small fluctuations of different parameters will no longer be visible). Therefore, at this stage it was decided not to use the Fourier transform, but use Dyubeshi Wavelets instead.



*Figure 1. Result of smoothing using Wavelet analysis*

**Anomalies fixing, filtering.** Figure 2 clearly shows abnormal points, where values significantly different from other points, or different from points of their class. Such abnormalities appear in data because of:

- Impaired recording procedure;
- Specific error occurred while recording induction log (IK). The anomaly was recorded due to the presence of a metal ring in the filter construction. It arises in the case of logging of IK after well completion.



*Figure 2. Example of anomaly*

In presence of this anomaly, it is impossible to carry out normalization of data, since values of anomalous points are significantly different from an average level. There are several ways to define such type of anomalies such as wavelet analysis, modified version of the linear classification algorithm (KNN).

Wavelet analysis can compensate anomaly, but at the same time it cannot determine exactly where it is. Moreover, compensation described above is not applicable in our case, so search and compensation are divided into two separate subtasks.

Figure 3 shows the algorithms developed in the module, which allows to successfully recognize anomaly objects.

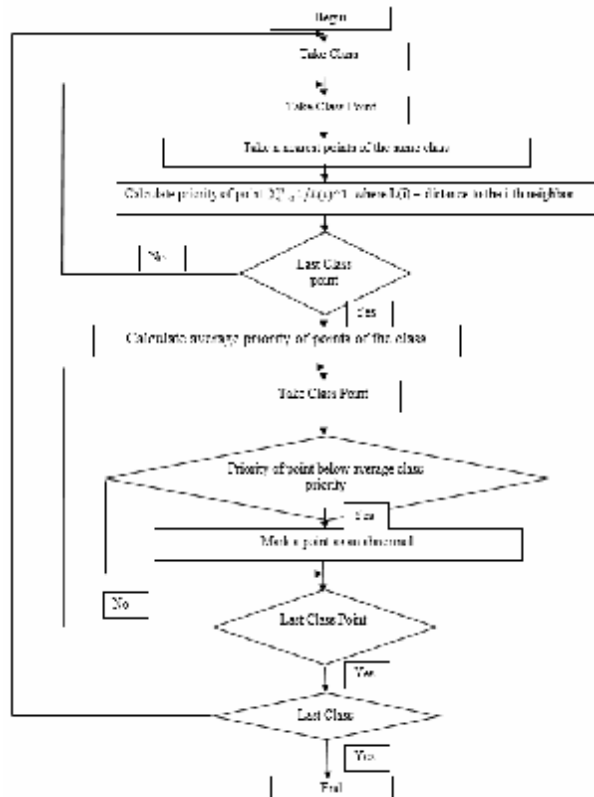


Figure 3. Modified KNN scheme to find anomalies

**Different recording level of signals** (Figure 4). The main reason of this anomaly – is acid seeping through lithological layers, during wells exploration. When this occurs PS recording level significantly increased, despite dispersion is virtually constant, so it is possible to restore level using the following algorithm (Figure 5).

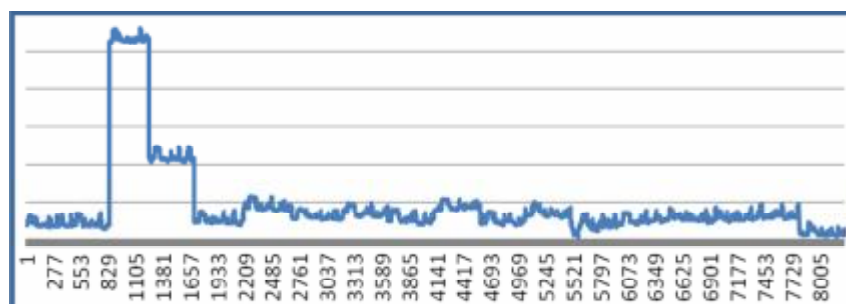


Figure 4. Different recording level of signals (data from several wells)



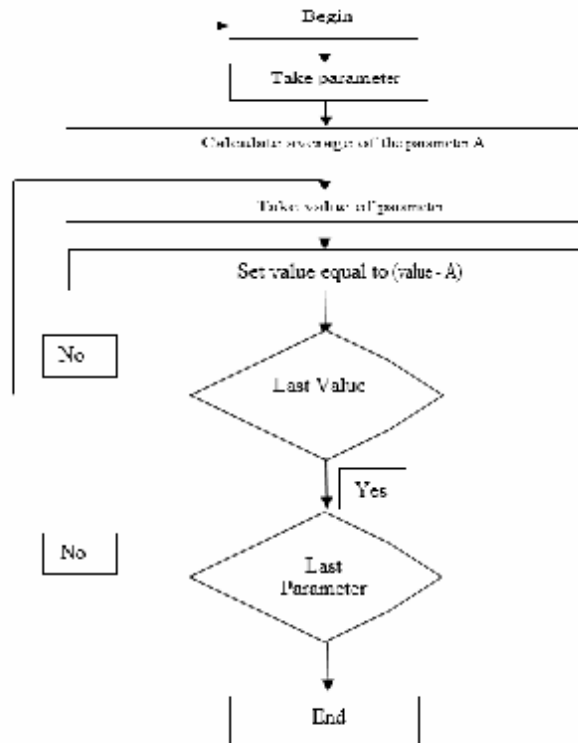


Figure 5. Algorithm to fix different recording level of signal

**PS Inverse.** KS (resistivity log) and PS (spontaneous polarization log) have an inverse relationship. PS can invert due to increased salinity of groundwaters.

*Algorithm:*

1. Search for an interval where KS log and PS log not in phase;
2. Connect start and end point by line;
3. Lay off new points symmetrically to line (Figure 6 a);
4. Connect consequent extrapolated points by lines (Figure 6 b);
5. Omit verticals from source points to lines from step 4;
6. Intersection points (white points) are result of recovery. (Figure 6 c).

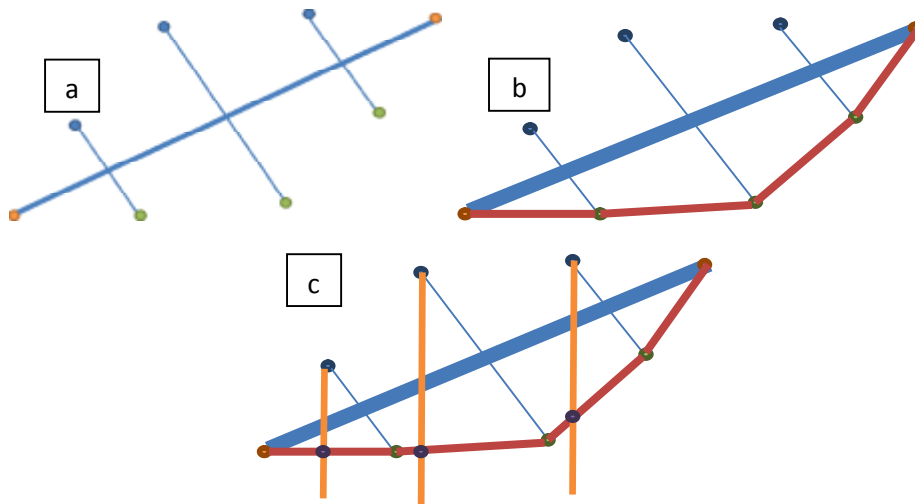


Figure 6. PS inverse fixing algorithm: a – 3rd step b – 4th step c – 6th step

Green points represent source data, blue points represent extrapolated points and white points represent recovered points.

**Removing log shift.** This anomaly (Figure 7) can occur while recording IK. Recording KS log and PS log conducted by one device, therefore there cannot be any shift between them, however IK log is recorded by separate device, so it is possible to have shift with other log data. Shift means that IK log begins from a depth different from a depth of other parameters.

IK and KS have an inverse relationship, since one describes conductivity (IK) and another resistance (KS). On a basis of this relationship shift can be identified and fixed by algorithm in Figure 8.

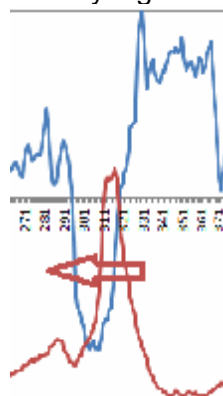


Figure 7. Log shift

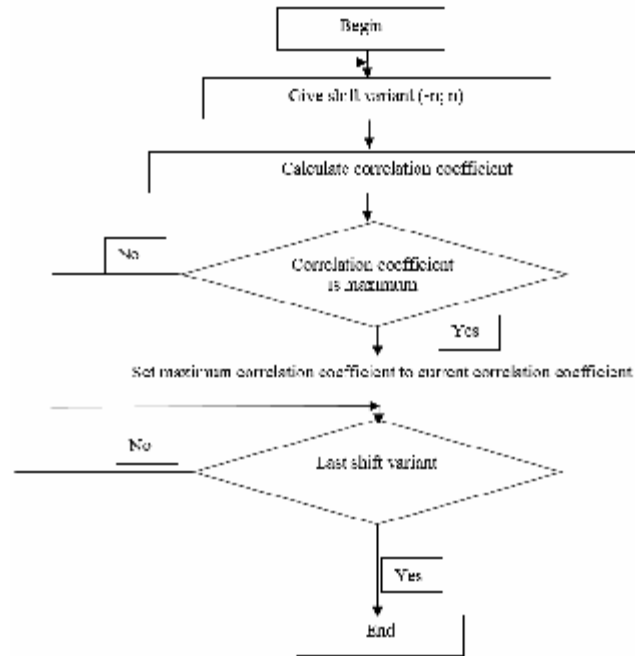


Figure 8. Algorithm to remove log shift

**Data normalization. linear normalization.** For each parameter linear normalization is conducted separately by the following formula (Formula 1):

$$X_i = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where  $X_i$  – current value;  
 $X_{min}$  – sample local minimum;  
 $X_{max}$  – sample local maximum.

Depth normalized by the same formula, but  $X_{min}$ ,  $X_{max}$  are global sample minimum and maximum respectively.

**Removing noise points.** As can be seen from Figure 2 there are points which located among points of a different class. Obviously correctness of such points interpretation is under a question. Such anomalies are called 'noise points', and they can destabilize and slow down NN training. These points may be identified by modified KNN method.

**Algorithm:** At first, priority of each point should be calculated. To do this, for each point  $n$  nearest neighbors are taken (distance is defined as

Euclidean from a given point). Then priority is calculated by the following formula (Formula. 2):

(2)

where  $\sum$  is the sum of point priorities among nearest neighbors of a same class as given point class;

— highest priority among defined class priorities.

Points with priority less than or close to 0 should be labeled as noisy, and can be removed.

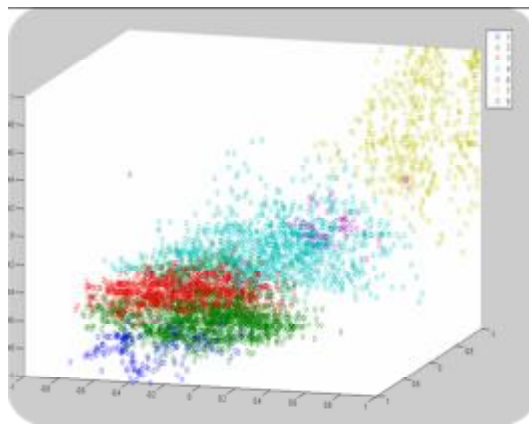


Figure 9. Result of removing noise points

The experiments on real data sets have shown the effectiveness of developed algorithms:

Well	Result
1	0.384
2	0.587
3	0.663
4	0.549
5	0.549
6	0.631
7	0.603
Average	0.567

Table 2. Results before preprocessing

Well	Result
1	0.593
2	0.625
3	0.631
4	0.574
5	0.571
6	0.662
7	0.590
Average	0.606

Table 3. Result after preprocessing

The tables above show the results of interpretation using metric method implemented in "Preprocessing Module" before and after preprocessing. Result - the share of correctly classified objects in comparison with the interpretation of an expert geologist.

Preprocessing not only does increase the average result, but also reduces the variance: example shows that interpretation result for well 1 was significantly different from the average before preprocessing, but after preprocessing well 1 result increased to an average value.

### **Conclusion**

Developed software "Preprocessing Module" implements a number of algorithms for data processing and data analysis tools. The experimental results showed the efficiency of the "Preprocessing Module" on real data sets – quality was improved and dispersion of recognition results was decreased. The developed functions successfully address the main problems with geophysical research data, such as:

- Noise in signals
- Anomalies in signals
- Different recording level of signals
- PS inverse
- Non-normalized data
- Noisy samples
- Shift of signals

One of the main topics for further research is issue of methodology for estimation of recognition quality and reevaluation of the developed algorithms using new quality methodology.

This work is performed by supporting grant №2318/GF3 of the Ministry of Education and Science of the Republic of Kazakhstan

### **References**

- [1] Ravil I. Muhamediyev, Yan Kuchin, Elena Muhamedyeva (2012). Recognition of Geological Rocks At the Bedded-infiltration Uranium Fields by Using Neural Networks. 2012 IEEE Conference on Open Systems, Kuala Lumpur, IEEE Xplore database, <http://cs.ieceemalaysia.org/ICOS2012/> 6 p. (accepted 08.2012) ISBN: 978-1-4673-1044-4, Digital Object Identifier : 10.1109/ICOS.2012.6417622M. Tim Jones. Artificial Intelligence: A Systems Approach. INFINITY SCIENCE PRESS LLC Hingham, Massachusetts, New Delhi, 2008. ISBN: 978-0-9778582-3-1
- [2] Ravil Muhamedyev, Yan Kuchin, Elena Muhamedyeva, Pavel Gricenko, Zhusulan Nurushev, Kirill Yakunin. (2012) The analysis of the data of geophysical research of boreholes by means of artificial neural networks. Proceedings of the Fourth International Conference "Informatics in Scientific

Knowledge”(ISK’2012) organized by Varna Free university and Institute of Mathematics and Informatics of Bulgarian Academy of Science, Bulgaria. p. 198-206, ISSN 1313-4345

[3] E. Amirgaliev, Z. Isabaev, S. Iskakov, Y. Kuchin, R. Muhamediyev, E. Muhamedyeva, K. Yakunin. Recognition of rocks at uranium deposits by using a few methods of machine learning Proceeding of The 14th International Symposium on Advanced Intelligent Systems November 13-16.2013 in Daejeon, Korea, [www.isis2013.org](http://www.isis2013.org), pp. 735-744, ISSN : 1738-8073

[4] Амиргалиев Е. Н., Исаков С. Х., Кучин Я. В., Мухамедиев Р. И. Методы машинного обучения в задачах распознавания пород на урановых месторождениях. //Известия НАН РК , 2013, №3. С.82-88.

[5] Kostikov D.V. Tools of interpretation of geophysical researches of boreholes on the basis of the transformed logging diagrams by means of a multilayered neural network.-M.: Russian state library, 2007.- 189 p.).

[6] Амиргалиев Е. Н., Исаков С. Х., Кучин Я. В., Мухамедиев Р. И., Уалиева И. М., Мухамедиева Е. Л. Распознавание пород на урановых месторождениях с использованием методов машинного обучения. Совместный выпуск: Вестник восточно-Казахстанского государственного университета/Вычислительные технологии.Институт вычислительных технологий Сибирского отделения РАН. Том «Информационные и телекоммуникационные технологии».- Усть-Каменогорск, ВКГТУ им. Д. Серикбаева, ИБТ, 2013. ISSN 1561-4212, 1560-7534. С.232-240.

**Authors:**

**P. Gricenko**, student-researcher, International IT University, Almaty

**Y.I. Kuchin**, senior engineer “GeoTechnoService”, Almaty

**R.I. Muhamedyev**, professor, head of chair, IT department, International IT University, Almaty

**Z. Nurushev**, student-researcher, International IT University, Almaty

**K. Yakunin**, student-researcher, International IT University, Almaty

**Contacts:**

Ravil Muhamedyev, email: [ravil.muhamedyev@gmail.com](mailto:ravil.muhamedyev@gmail.com)

phones: +7 701 3393720, +7 777 2399788

## **INFORMATICS IN EDUCATION**

### **PROGRAMMABLE TOYS AND SOFTWARE FOR SIMULATION WORK OF ROBOTIC SYSTEMS AS A TOOL FOR DEVELOPMENT OF COMPUTER SKILLS**

Angel Angelov

**Abstract:** *This report considers the possibilities for the development of computational skills in students through the use of embedded devices, robotic systems and simulation of their work in education.*

**Keywords:** *programmable toys, computer skills, computing skills, games, programming.*

Programmable toys and software for simulation work of robotic systems have found their place for several years now in the implementation of educational programs in natural and mathematical sciences at elementary, primary and secondary levels of Bulgarian education.

The use of embedded devices (hardware with is embedded software) [3], robotic systems and simulation of their work in education contribute to the learning of key concepts, such as:

- *orientation in terms of space* - which includes the assessment of distance, size, shape, relative position of objects and their position depending on the the milestone;
- *shape* - mediated generalized reflection of the essential properties of reality's mathematical objects
- *position and movement*
- *angle and others.*

Examples of the use of programmable devices and means of programming currently available in Bulgarian schools are:

- SunSpot & Sun Spot Java Development Kit
- Gadgeteer & C #
- Moway Robot & Scratch, Moway World, C, Assembler
- Raspberry PI & Python, Scratch
- Lego WeDo & Scratch
- Valiant Roamer & keys on the device
- Bee-Bot, Constructa-Bot, Pro-Bot & keys on the device
- Pixie keys on the device
- Finch & Javascript, C, C + +, VisualBasic

Programmable devices all share the same work principles – requirement to develop and implement an algorithm that is implemented by the device as a series of accepted commands and executed actions. Actions of the devices can be the following:

- from a motion in a certain direction or following a straight line to a complex pattern of movement as a study of the maze and finding the most direct route;
- from simulation of the screen to the orientation among objects using volumetric, infrared and other sensors ...

The wide range of programmable educational resources and the opportunity to be managed by different logical programming tools not only make them possible to use by different age groups - from preschool to secondary school level in the Bulgarian education system, but they also offer a new alternative way of developing skills in logical thinking, programming skills, computing skills, etc..

Since the development of these skills is directly linked to the development of multiple intelligence, computing skills and programming skills, they need to be considered as part of the logical-mathematical intelligence. In turn, logical-mathematical intelligence involves the development of logic, abstractions, justifications and problems dealing with numbers and symbols.

Computing skills are the focus of several studies and reports in recent years. Guzdial [4], Qualls & Sherrell [64] and Wing [7] define these skills as a set of intellectual and thinking skills that enable an individual to operate successfully with numbers and symbols for their association, which includes the use of methods and programming languages in order to solve problems and tasks.

It is these opportunities in the development of computational skills contribute to the successful integration of operations with programmable educational resources in curricula, and from there into teaching units in different subjects.

Using programmable educational resources in the learning process can be divided into two areas: the use of these resources for educational purposes and their programming in order to achieve a particular result [5]. The latest involves the constructivist approach, combined with the very active approach to training and research approaches which can be applied

Although literature lacks empirical evidence to support the development of computational skills using programmable toys and simulation software for the operation of robotic systems, a number of authors [1], [2], [7] believe that they help very successfully develop logical thinking, algorithm development, debugging, and also address issues and problems from a different angle:

- how to resolve a problem / task in the quickest way
- how to resolve it in the most optimal (best) way



- how to reformulate the problem in order to move towards a known example

- how complex is the task
- how to resolve the problem
  - thanks to Computer Technology
  - using purely theoretical sciences

how to interpret the task / problem - code data and data source

In order to develop computational skills using programmable educational resources, students need to be put in situations that require them:

- establish and implement algorithms for a particular problem;
- be able to evaluate the algorithm on certain criteria
- use different calculation methods with the help of computer technology and purely theoretically
- detect and correct logic errors and errors in algorithms and others.

The inclusion of robotics in education supports the development of students in the field of natural and mathematical sciences, develops computing skills and programming skills.

On the other hand, in order to establish and avoid logic errors and algorithms used in the management of programmable educational resources, students need to simulate and compare the work of several algorithms. Computational skills are of utmost importance here and are closely related to the competencies of the students in terms of the use of different computing methods, computer technologies and theories.

As a conclusion, it could be highlighted that the inclusion of robotics (programmable toys and software for simulation work of robotic systems) in education supports the students' development in the field of natural and mathematical sciences, develops their computational skills and programming skills.

## References

- [1] Ater-Kranov, A., Bryant, R., Orr, G., Wallace, S., & Zhang, M. (2010). Developing a community definition and teaching modules for computational thinking: accomplishments and challenges. Paper presented at the Proceedings of the 2010 ACM conference on Information technology education.
- [2] Barnes, T., Richter, H., Powell, E., Chaffin, A., & Godwin, A. (2007). Game2Learn: building CS1 learning games for retention. SIGCSE Bull., 39(3), 121-125.
- [3] Goel, Anish, Introduction to Embedded Systems, (<http://goo.gl/z11H2>), 10.7.2013
- [4] Guzdial, M, Education: Paving the way for computational thinking. Commun. ACM, 51(8), 25-27.

[5] Momcheva Galina, Angelov Angelo, Programmable toys for pre school and primary school age

[6] Qualls and Sherrell, Why computational thinking should be integrated into the curriculum. J. Comput. Small Coll., 25(5), 66-71.

[7] Wing, J.M., Computational thinking. Communications of the ACM, 49(2), 33-35.

**Author:**

**Angel Stoyanov Angelov**, Secondary School "Sava Dobroplodni", Shumen, +359 899359699, [angelov@innovateconsult.net](mailto:angelov@innovateconsult.net)

## APPLICATION OF GAMIFICATION IN UNIVERSITY MARKETING

Teodora Bakardjieva, Milen Sotirov

**Abstract:** *In the present paper some gamification mechanics concerning university marketing are discussed. The application of gamification to non-game systems is much more appealing, interesting and fun for the auditorium. Some gamification strategies and game mechanics are implemented at Varna Free University to solve problems and create more engaged students and visitors. It can be seen how mechanics can help drive very strong community engagements, as well as how slight changes in mechanics can yield very different behaviors.*

**Keywords:** *gamification, events marketing, educational games, social networks*

### Introduction

Gamification has been a trending topic [1][2][3][4] and a subject to much hype as a means of supporting user engagement and enhancing positive patterns in service use, such as increasing user activity, social interaction, or quality and productivity of actions [5]. Nowadays gamification is considered as the next generation method for marketing. [6] The beginning of that trend was in 2011 and Gartner Group forecasted that 70% of top 2,000 companies in the world would utilize applications with the concept of Gamification applied in 2014 and more than 50% of organizations would gamify their innovation processes. [7][8] There are a lot of theories in which the assumption on which game theory is based is too restrictive and is too theoretic to be widely employed in managerial practice [9][23][24][25]. Furthermore, game theory by Von Neuman and Morgenstern (1944), research has been debating about the possibility to apply game theory to solve marketing problems, and in particular to use it as a tool to predict competitive behavior [26]. Later on the debate has been extended to all the other possible uses of game theory in marketing. Management decisions about marketing mix have to be taken in situation of competition and variability in the business environment. Porter [27] clearly states the relevance to consider the possible effects of competitor's strategic decisions on every level of managerial decision process.

The application of gamification to non-game systems is much more appealing, interesting and fun for the auditorium, no matter if we talk here about customers, actual or potential customers, suppliers, distributors etc. The issue is how marketing needs could be satisfied using attractive way of sending the information, having an interesting and catchy headline or title.

## **Background**

Gamification makes joy of game such as points, level-up, ranking, challenge tasks, and compensation. According to Entertainment Software Association (ESA), 72% of households in America enjoy games. In 2011, those who are over 50 account for 29% of the gaming population, which shows that game is not for children [10]

More and more people in various age slots are enjoying games in leisure time and game becomes more than a play of entertainment. Gartner Group forecasted that 70% of top 2,000 companies in the world would utilize applications with the concept of Gamification applied in 2014 and more than 50% of organizations would gamify their innovation processes [11] [12].

University marketing and gamification are very close because the students' audience is open to system of competitions and rewards. Zichermann & Cunningham (2011) present a gamification approach called SAPS, which stands for: Status - is the relative position of an individual in relation to his peers or strangers, in a defined ranking system. In games badges or levels can be awarded. Access - for the audience, in the point of view of the events, can be the main issue if they are going to take part, or not, first of all depending on the location and the time when the event occurs. From the game point of view access can be granted very easily if the games are deployed via the internet, which is worldwide, the only restraint being a good connection. Organizations can reward the most deserving players by rewarding them with VIP treatment, or access to important meetings, or even allowing early and preferential discounts. Power - which can be relatively given as long as the game designers allow it, is considered to be a great success, also as an honor for players provided they take their roles very seriously, for example, granting them moderator or master features and in this way they get greater power over the others players. Stuff - like giveaways, as gadgets, that are unique and produced in limited series, especially for the the game, can have a great value and be highly appreciated by the participants [13].

Gamification approaches have to be used before the event to arise the interest and, at the same time, to be as a surprise for them, that their effort will be recognized and appreciated. So the player and the participant will become the audience.

## **Discussion**

Elements on the university websites:

Table 1. Game elements

Human Needs	Game Mechanics
Status	Levels
Competition	Leaderboards
Achievement	Challenges
Reward	Points

## Top 10 Mechanics for University Marketing

### 1. Achievements - Badges, Ranks, Points and Levels

In addition to signaling status, students desire badges, points, ranks and levels for all kinds of reasons. For many students, collecting is a powerful drive. Other students enjoy the sudden rush of surprise or pleasure when an unexpected badge shows up in a gamified system. A well-designed, visually valuable badge can also be compelling for purely aesthetic reasons. For universities, badges, ranks and levels are an excellent way to encourage social promotion of the products and services. Ranks, Badges and Levels also mark the completion of goals and the steady progress of play within the system[6].

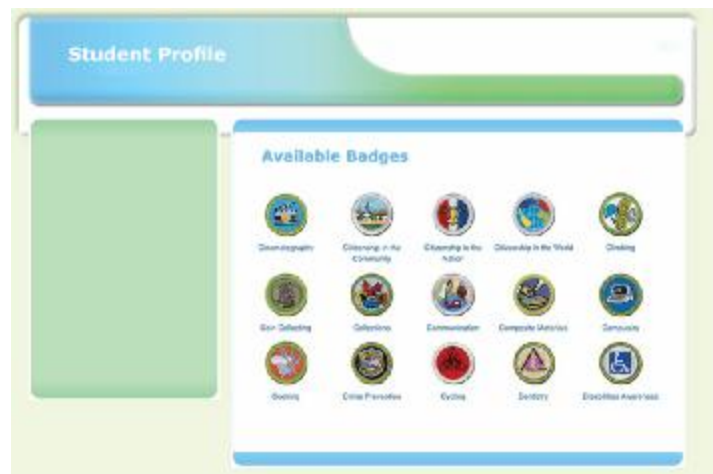


Fig. 1. Available Badges

### 2. Challenges and Missions

Challenges and Missions give students direction for what to do within the world of the gamified experience. After all, how much fun would a scavenger hunt be if the student were told to just go and find some stuff? Nothing specific...just stuff. There isn't much of a challenge there. There is zero intrigue and an absolute lack of structure. Some students enter the game with no idea of its goals or fundamental drives. So, even if

a challenge isn't at the front and center of the experience, using challenges as an option somewhere in the body of the system can add depth and meaning for the student[6]. Completing a challenge gives a student satisfaction from reaching a goal.



Fig. 2. Achievements Showcase

### 3. Integrated Profiles in Student Status

The profile shows student's current status and rank, earned badges, etc. and provides a complete overview of any given student stands [14]. From the profile all students can see their achievements, badges, points, levels and they can easily share all the information with the social media buttons integrated in their profile.

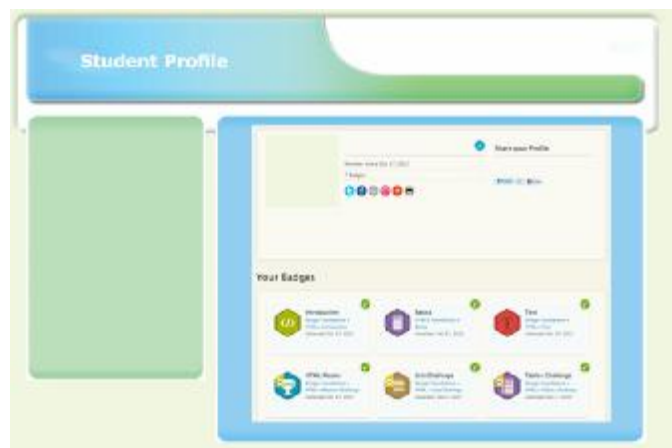


Fig. 3. Student Profile Overview

#### 4. Leaderboard Statistic and Activity Feed

The purpose of a leaderboard is to make simple comparisons. Unsurprisingly, most students don't need any explanation when they encounter a leaderboard. By default, the student see an ordered list with a score beside each name, and they understand that they are looking at a ranking system [6]. Making all visible to other students and providing leaderboard is a way to encourage students to compete.



Fig 4. Leaderboard

The Activity Feed shows what has been taking place in the university system overall. This shows the students that they are not alone and can often show a student a feature or piece of engagement they were previously unaware of [15].



Fig. 5. Activity Feed

## 5. Search for Hidden element

For example: the University can place hidden images and QR Codes etc. on the websites. Visitors and student have to search for and locate the QR code or the images etc. The student can also search for secret messages, secret #hashtags on various platforms such as websites and social media profiles. The first 5 visitors or students that found the elements will receive points, badges and rewards.



Fig. 6. Hidden QR Code

## 6. Unlocks

Every student and visitor can find a list of badges, levels and information on the website or in their profile, the list contains information such as unlock information, available badges etc.



Fig. 7. List of available badges and levels



## 7. Rewards and Discounts

Gamification techniques strive to leverage people's natural desires for competition, achievement, status, self-expression, altruism, and closure [16]. One of the main gamify strategy is the rewards for students and visitors who accomplish Goals, Challenges and achievement, they are awarded with Points, Badges, Levels, Certificates, Master Class, Discounts and etc.



Fig. 8. Rewards& Prizes

## 8. Real-World Activity

Real-World activities helps the students earn points and badges. Creating a fun competition between students for example to take a picture on the university campus and posting it on the website or in social media channels. The picture with the most votes and shares wins and the owner earn points and badges. That's one method to create social media buzz. There can be Daily, Weekly and Monthly Contests.



Fig. 9. Photo Contest

### 9. Social Sharing - Creating Buzz

One way to get people buzzing about a brand over social mediachannels is to "gamify" the experience and engaging students in fun ways. The University can drive repeat traffic to the social media by creating a contest or through longer-term projects that tie into the features of specific social platforms, such as Facebook and Google [17].For example students win reward with weekly prizes by completing various challenges like quizzes, questions and finding things on the web over a period of time [21].Students get involved by writing reviews and engaging with other students to create buzz surrounding the University and they can earn points for their activities and status.



Fig 10. Social Buzz leaderboard

### 10. Give Students Voice & Choice

In many games, students are often given various choices that affect their progress, and the outcome of the game. Some are rewarded with extra points or badges, others benefit when completing harder challenges. The same choices can be implemented into gamified websites, allowing students different options they can use to reach their academic goals [18].

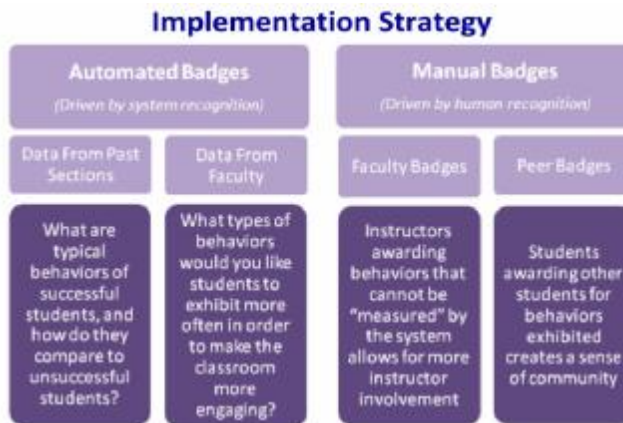


Fig. 11. Implementation Strategy [19]



Fig. 12. Statistic [19]



Fig. 13. Player Centered Design[20]

## **Results**

Some gamification strategies and game mechanics are implemented at Varna Free University to solve problems and create more engaged students and visitors. It can be seen how mechanics can help drive very strong community engagements, as well as how slight changes in mechanics can yield very different behaviors. The students' desire to advance in leaderboards, levels and to receive visual acknowledgment of their progress (badges, ranks, etc.) propels them to engage further in university activities. The real-time activity stream that students can view to compare their achievements against other student's achievements infuses a healthy amount of competition into higher education. The ambition to be a top ranking student is heightened by the use of gamification strategies, which in-turn increases the amount of work students are willing to do. By driving additional engagement the University has taken the learning experience to the next level, boosting motivation, participation, and academic success.

## **Conclusions**

Once you've got student's attention, we can amplify it using the power of social connection, progression to mastery, and viral loops. These gamified dynamics must be constantly tested, measured, and refined to optimize the experience for what the student needs. And it's critical to never let it get boring; even the best games can lose steam if you don't keep them fresh. If you are successful at maintaining engagement with gamification, you can parlay that into a new-style loyalty program that minimizes tangible rewards, instead paying out in status, access, and power. The best examples have even managed to monetize this concept, turning loyalty points into a profit center and jousting for dominance in a strategic global market for virtual currency. Students who are engaged and aligned with your brand vision can do even more. With the right incentives, you can get the crowd to design products, solve problems, and create intellectual property that you can monetize. This concept, called crowdsourcing, has yielded major solutions to science, economic, and creative problems. This proves that you can even get lower-skilled customers to do hard work and solve real problems without having to shy away from the tough stuff. If everyone is aligned, properly incentivized, and enjoying themselves, the sky's the limit.

## References

- [1] Kapp, Karl. (May 1, 2012) The Gamification of Learning and Instruction: Game-Based Methods and Strategies for Training and Education. San Francisco: Pfeiffer. ISBN: 978-1-118-09634-5. 336 pages.
- [2] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: defining gamification", In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, September 28-30, 2011, Tampere, Finland, ACM, pp. 9-15
- [3] J. Hamari, and V. Lehdonvirta, "Game design as marketing: How game mechanics create demand for virtual goods", International journal of business science and applied management, 5(1), 2010, pp. 14-29
- [4] K. Huotari and J. Hamari. "Defining gamification: a service marketing perspective", In Proceedings of the 16th International Academic MindTrek Conference, October 3-5, 2012, Tampere, Finland, ACM, pp. 17-22.
- [5] J. Hamari, "Transforming Homo Economicus into Homo Ludens: A Field Experiment on Gamification in a Utilitarian Peer-To-Peer Trading Service", Electronic Commerce Research and Applications, 12(4), 2013, pp. 236- 245)
- [6] Zichermann, Cunningham, "Gamification by Design", O'Reilly Media, 2011.
- [7] Yi, Ju-HoA " Study on applying the concept of Gamification into Architecture Planning"((Proceeding of Annual Conference of the Architectural Institute of Korea) vol.32, pp.179-183, 2012.
- [8] Jon Radoff, park gi sung Trans, "Game on : energize your business with social media games ", Acorn, 2012. Jon Radoff, 2012
- [9] Wagner H. M. Principles of Operations Research. NJ (USA) Prentice Hall, 1985.
- [10] Jong-UkHeo , "Analysis of Web Visitor Behavior using Public Data Analysis and Gamification", (Human Computer Interaction Korea ), pp. 239-242, 2013)
- [11] Yi, Ju-HoA " Study on applying the concept of Gamification into Architecture Planning"((Proceeding of Annual Conference of the Architectural Institute of Korea) vol.32, pp.179-183, 2012
- [12] Jon Radoff, park gi sung Trans, "Game on : energize your business with social media games ", Acorn, 2012
- [13] Moise, Daniel and Cruceru, Anca., The Use of Gamification in Events Marketing, International Journal of Economic Practices and Theories, Vol. 4, No. 2, 2014, Special issue on Marketing andBusiness Development, e-ISSN 2247-7225
- [14] Gigya INC. <http://www.gigya.com/>
- [15] Badgeville <http://Badgeville.com>
- [16] Big Fish Games <http://www.bigfishgames.com/blog/what-is-gamification/>
- [17] Entrepreneur Media, Inc. <http://www.entrepreneur.com/video/224177>
- [18] TeachThought <http://www.teachthought.com/teaching/how-to-gamify-your-classroom/>
- [19] School of Business Switzerland <http://web.fhnw.ch/plattformen/blogs/gamifyu/2013/07/05/gamified-education/>

- [20] UX Magazine <http://uxmag.com/articles/five-steps-to-enterprise-gamification>
- [21] <http://www.gamification.co/2013/07/18/kaplan-gamification-system-shows-more-studentsengagement/>
- [23] Lazer W. e Thomas G. D. Handbook of Market Research. New York: McGraw-Hill, 1974.
- [24] Moorthy K. S. Using Game Theory to Model Competition. Journal of Marketing Research, 1985, 23: 262-282.
- [25] Tullock G. Adam Smith and the Prisoner's Dilemma. Quarterly Journal of Economics, 1987, May, p. 179-221.
- [26] Herbig P.A. Game Theory in Marketing: Applications, Uses and Limits. Journal of Marketing Management, 1991, 7: 285-298.
- [27] Porter M. E. Competitive Strategy. New York: Free Press, 1980.

**Authors:**

**TeodoraBakardjieva** Prof., PhD, Varna Free University,  
Director Technology Institute

Phone: 052/359-524, email: [bakardjieva@vfu.bg](mailto:bakardjieva@vfu.bg)

**MilenSotirov** Assist. Prof., ExpertSEO/SEM, Varna Free University,  
Phone: 052/359-537, email: [msotirov@vfu.bg](mailto:msotirov@vfu.bg)

## MODEL FOR APPLICATION OF GAME-BASED LEARNING IN SECONDARY SCHOOL

Todorka Glushkova

**Abstract:** Educational games become an increasingly important component in the modern learning process. The article presents a model for development and use of games in classroom training and in the self-training of the students from Secondary school "Hristo Smirnenski", Brezovo. Development of educational games from students that are applied in the learning process for all classes and forms of training increases the interactivity, motivation and effectiveness of the all learning process.

**Keywords:** educational games, interactivity, e-Learning

The modern society places higher requirements and criteria for education. Today is becoming increasingly important to teach students how to extract and process the information. They need to acquire competence in teamwork in the development and implementation of projects. We have to teach them to be active subjects in their learning. In this sense a special place in the learning process take ICTs and the interactivity, which these technologies provide [3]. This article will look at some aspects of interactivity and opportunities for improving the effectiveness of learning through application of game-based approaches [2].

The teachers and students from secondary school »Hristo Smirnenski« Brezovo use opportunities of the SCORM-based school educational portal that provides different levels of access of its users, according to their profile characteristics [7],[8]. During the current academic year students from the School of IT decided to create several educational games to help their classmates. These games are provided as separate services in their respective learning environments in the portal, and can be used in classroom and for independent work from all students.

According to the used technologies, developed games are various types as follows:

- P** By application of the MS.NET – technology;
- P** By using of J2EE technology;
- P** By using of SCRATCH-based programming tools.

### **Some games from the first type:**

- "Adventure in Znamlandiya " is a game, written in Visual Basic. It is a multiplayer game designed for self-study of fourth graders for their preparation to National Exams. Students play in the local school network and compete with each other for the title "King of Znamlandiya" They must address the various issues and pass through many trials related to

teaching of main school subjects (Fig. 1). As a result, students showed better results at the end of the school year.

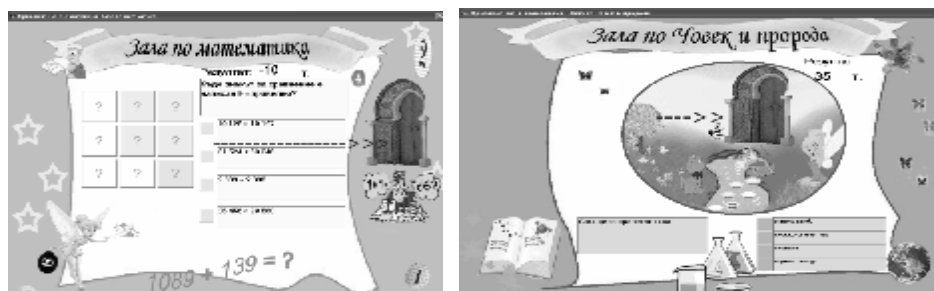


Fig. 1. "Adventure in Znamlandiya"

• **«Adventure in the Ice Kingdom»** is a learning game developed on Visual Basic too. It is designed for training of students in the third grade. The game is designed for the local network of the school. The administrator sets the work of the different levels and user environments; the teachers create and write questions and tasks for different levels, students learn by playing in the enchanting story of the Ice Queen Kingdom. The aim is to release the fairy tale characters trapped in the ice castle of the evil Queen. The children have to pass through various tests and they have to answer the randomly generated questions from different school subjects. (Fig.2.) The game used for training of the pupils in the afternoon classes.



Fig. 2. „Adventure in the Ice Kingdom"

• **“Smyatanitsa”, “Bukvenka”, “Game for top achievers”** and others games are developed in C#. These games are designed for pupils in accommodation centers for children deprived of parental care, as well



as other students with knowledge gaps of different classes. (Figure 3.) They are desktop applications and are available in a specially developed website "Home for our children" from educational portal of SOU "Hristo Smirnensky" Brezovo, section "Services" (on web-address <http://sou-brezovo.org>)

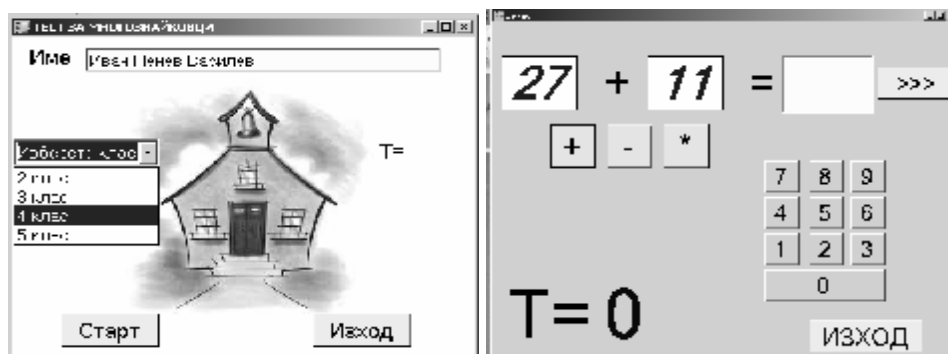


Fig.3. C# educational games

#### Some games from the second type:

- **"Discover Bulgaria ... and stay here!"** is an WEB application for preparation of the Government Exams in geography, biology and history. The site was created by using of the J2EE technology with server part - JSP and servlets and database on MySQL server. Except many references and general information, the site also offers an online game in which graduates can check your level of preparation for Exams. Several times during the year, together with the author Miroslav Savov, we organized competitions with prizes for the top students-gamers, which adds an additional adversarial element. As a result, the average grade of the Exams in these subjects is almost half a unit higher than in previous years. (Fig.4.)



Fig.4. „Discover Bulgaria ... and stay here!"

- **„Logic”** – is an WEB application for logical games, puzzles, IQ-tests and tasks. It is developed in Java by using of J2EE technology. The site also offers an interesting logical game with riddles, puzzles, enigmas and more, which further encourage students and increases their activity. Every week we announce a competition for the most complete solution to the some logical problem or enigma, and the student with the best solution receives award from the school leadership. All of this stimulates largely activity, motivation and logical thinking of students. (Fig. 5).

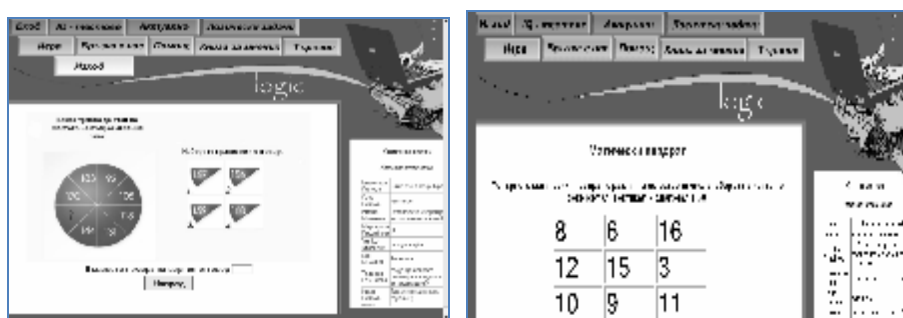


Fig. 5. “Logic” game

- **„INSULA EUROPAE”** – is a learning game for subjects in social sciences and is used in the teaching of topics related to European integration, European institutions, European literature and culture. The game was created on the initiative of school clubs "Insulza Europe" on a joint international project with partners from nine European countries. By virtual seminars all students defined the architecture and functionality of the game. Our students from the School of IT wrote the code, all others students defined, selected and sent questions. We announced a general online competition, the results of which announced on May 9th. (Figure 6.). This game was also developed based on J2EE technology using Java- servlets, JSP and MySQL DB server.

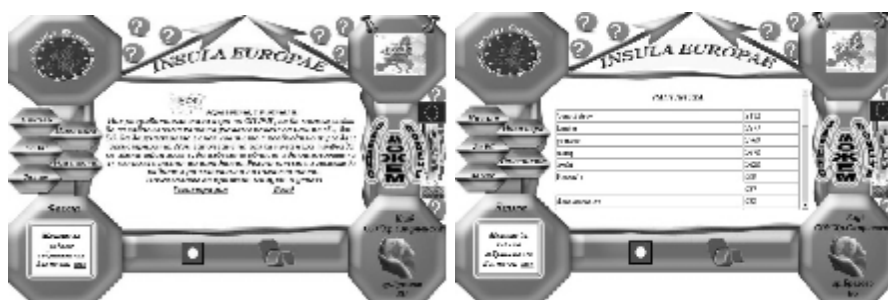


Fig. 6. „INSULA EUROPAE” game

- **“At school as a game”** is a web application designed for online preparation of fourth grade for national external evaluation. Except the game there is helpful information on the inspected school subjects. The game has a competitive nature. Students answer questions with different levels of difficulty in their chosen subject. The students from top ten of the rankings receive prizes periodical. The application offers a separate environments for administrators and teachers, so questions and content in the database is continually updated.



Fig. 7. „At school as a game”

These and other web-based games are available from the school portal, section «Services» (<http://sou-brezovo.org>)

#### Some games from the third type:

- **“For mathematicians and not only”** and etc. are learning games for mobile devices - GSM, tablets etc. with operating system Android, designed for fun exercise of fundamental arithmetic operations from pupils at an early stage. They are developed in the programming environment of MIT App Inventor 2 [11] (Figure 8) The creation of mobile applications such these increased interest of students in the School of IT and all the students who use them.

Programming by using of SCRATCH-based [6], [10] languages and the use of the visual, object-oriented and component-based programming environment are useful for both initial familiarization with main structures and basic algorithms as well as for more advanced students.

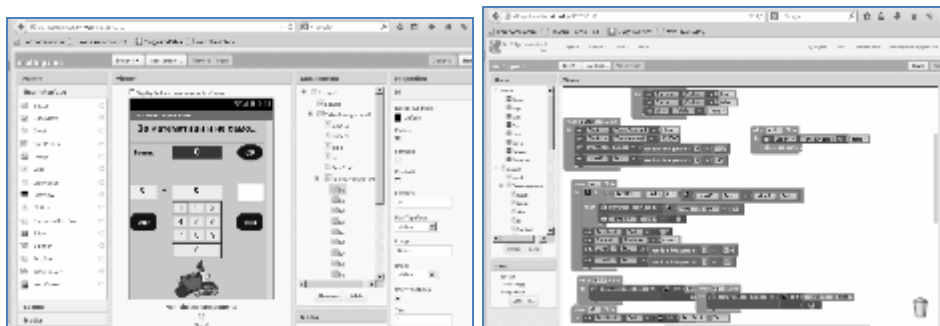
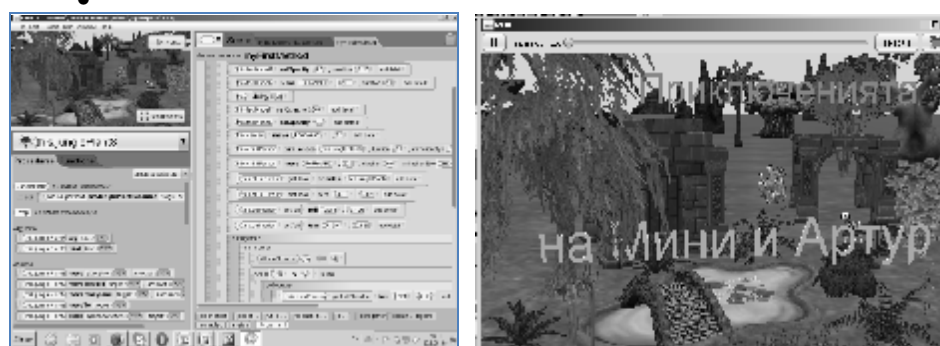


Fig. 8. Development of games for mobile devices by App Inventor

• **Developing of 3D games in Alice 3.1 programming environment** In the last month the students of the School of IT apply their knowledge for SCRACH based languages by developing 3D games in programming environment Alice 3.1. [9] (Fig. 9). Although not yet ready full realization, the interest to work in a 3D environment is great.



Фиг. 9. Работа в среда Alice 3.1

In specialized literature there are known three levels of interactivity in teaching using ICT: the level of production, the level of selection and level of cooperation [4],[5]. These three levels of interactivity supported six categories in varying degrees associated with them - feedback and control, productivity, creativity, communication, adaptability. Students pass all these levels and categories by using of these games. They interact with the learning environment and get answers and assessments thereof; they receive assignments, make plans and apply different strategies; communicate and compete with each other.

On the other hand in the training of students in the School of IT I use another technique-playing role-play "software company" [1]. We form

different teams as: graphic designers, C # and MS.NET programmers, web developers etc. The members of this "software company" define project manager for each game, create plan of work. The team of programmers writes source code, they make test of the products and present them to users. In combination with other interactive methods and services of the educational portal, the students from the IT School create educational games as they themselves play game. This increases their motivation and activity. As a result, some from created IT products were successfully presented in national Olympiads, competitions and conferences on IT.

In conclusion we can say that interactive methods and game-based techniques in teaching various school subjects increases motivation and willingness of students as well as the effectiveness of the learning process. All games and other training services are available from the section «Services» of the educational portal of SOU "Hristo Smirnenski" on web-address <http://sou-brezovo.org>.

### References

- [1] Glushkova, T., Pupil's research school on IT, book "Science education in European schools", pp.8-10, Italy, [http://www.stella-science.eu/documents/STELLA\\_eBook.pdf](http://www.stella-science.eu/documents/STELLA_eBook.pdf)
- [2] Prensky, Marc, Digital game-based learning, Computers in Entertainment (CIE) - Theoretical and Practical Computer Applications in Entertainment, Volume 1 Issue 1, October 2003, EISSN: 1544-3574
- [3] Momcheva G., Veselina Spasova, GAMES AS A CHALLENGE IN EDUCATION, The 10th INTERNATIONAL CONFERENCE INFORMATION TECHNOLOGIES AND MANAGEMENT 2012, April 12-13, 2012, Information Systems Management Institute, Riga, Latvia
- [4] Shedroff, N., Information Interaction Design: A Unified Field theory of Design. In Jacobsen, R (Ed.), Information Design, pp. 267-292. Cambridge, MA: MIT Press, 1999.
- [5] Stoyanova, A., T.Glushkova, Model for implementation of interactive distance learning, in proc. of International Science Conference "Informatics in the Science Knowledge", Varna, 24-26 June 2010, Varna, pp.405-418, ISSN 1313-4345, 2010 (и на CD).
- [6] Veselina Spasova, G. Momcheva, "Scratch games for preschool students", International Scientific Conference Informatics in the Scientific Knowledge, June 27-29 2012, Varna, Bulgaria
- [7] Глушкова, Т., Д. Плачкова. *Ученическата изследователска школа по ИТ*, В сб. „Иновативни методи и практики в българското училище“, София, 47-51, ISBN 978-954-92045-7-8, 2009., [http://www.old.cei-bg.org/sites/all/files/u1/inovativni\\_metodi.pdf](http://www.old.cei-bg.org/sites/all/files/u1/inovativni_metodi.pdf)
- [8] Глушкова, Т., Д. Плачкова. Интерактивно компютър-базирано обучение в СОУ "Христо Смирненски" гр. Брезово. В сб. „Иновативни методи и

практики в българското училище”, София, 57-60, ISBN 978-954-92045-7-8, 2009., [http://www.wold.cei-bg.org/sites/all/files/u1/inovativni\\_metodi.pdf](http://www.wold.cei-bg.org/sites/all/files/u1/inovativni_metodi.pdf)

[9] <http://www.alice.org/index.php>, visited on 30 May 2014

[10] <http://scratch.mit.edu/>, visited on 30 May 2014

[11] [ai2.appinventor.mit.edu](http://ai2.appinventor.mit.edu), visited on 29 May 2014

**Author:**

**Todorka Atanasova Glushkova**, PhD, Department of Computer Technologies, University of Plovdiv “Paisii Hilendarski”, Plovdiv and Secondary school “Hristo Smirnenski”, Brezovo, Bulgaria, 0878212950  
[todorka.glushkova@gmail.com](mailto:todorka.glushkova@gmail.com)

## ONLINE DOCUMENT COLLABORATION TOOLS IN EDUCATION

Antonina Ivanova

**Abstract:** *The paper highlights online collaborating tools for learning content creating which can be implemented in the educational process at Varna Free University. Online collaboration enhances learning process and increases student's engagement. Working together online and synchronizing documents on different devices helps students and teachers collaborate more effectively.*

**Keywords:** *online collaboration, cloud services, google docs, zoho, onedrive*

Collaborative educational technologies can enhance active participation (through content creation), increase students' engagement with course content, and enrich the learning process. The online collaboration has its pedagogical advantages like:[6]

- Encourage peer-instruction.
- Build a learning community
- Give students a sense of their learning level
- Allow efficient time management

The traditional way of collaborating to create a document is for one person to create an initial document and then email it out as an attachment to all other team members. Different revisions are sent back and work is required to put it together, since there are many different versions to be kept track of. The online collaboration alternative is to send an email containing a link to the document instead of the document itself. The main benefits are that each team member can modify the current, up-to-the-second version, and all revisions are tracked.

Online collaborating tools for creation learning content have mostly word-processor documents, spreadsheets, presentations, forms. All kinds of documents can easily be shared with anyone who has an account for the platform. Other people can be added as collaborators (with editing rights) or as viewers (without editing rights). Documents can also be downloaded in a number of formats. The most popular tools have options to publish documents as a web page for all to view.

In this paper three platforms for online collaboration are presented – Google, Zoho, OneDrive. Every one of them can be implemented in teaching and learning process. At Varna Free University students are familiar with collaborating on Google Drive. But for the variety of programs at the university the other platforms could be suitable too. That is the reason to organize different seminars for teachers to observe the functionalities of the platforms and to conclude possibilities for implementing online collaboration tools at the university.

For educational purposes some functionalities of the collaboration tools are welcome and when selecting the tool it is good to be available:

- Control edit settings: teachers can control the access level – edit, view, comment
- Allow simultaneous work – students and teachers can see who else is editing the document at the same time. The changes have to be immediately updated on the document and changes made by individuals are differentiated (for example different colors)
- Saving changes and retrieving past versions. The tool has to keep tracks of the changes and saves a history of different versions. If anyone delete a part of a document, to be possible to revert to previous versions.
- Online tutorials and help. This will allow students to easily clarify issues and problems without teacher's help

### **Google Drive / Google Docs**

Google Drive / Docs makes creating, editing and sharing documents, spreadsheets and presentations simple and free. The docs are stored safely online. It is possible to create a document from scratch, upload an existing document or browse the template gallery for a format that suits user's needs. Google Docs can be used to create an online form to collect data from others or for short quizzes. And it's easy to export or download the docs in a variety of formats.



Figure 1. Collaborating on Google Drive

Collaborating on group projects with colleagues can be a great way to boost someone's professional development. However, when several people are editing the same document simultaneously, this can sometimes lead to the existence of multiple drafts of the same file, which can result in confusion! Tools such as Google Drive are designed to enable to share information and documents more effectively and



efficiently when working with others on a joint project. One of the main purposes of Google Drive (formerly Google Docs) is to allow multiple people to edit the same document, spreadsheet, presentation, drawing, or form without creating duplicate copies. Documents can either be uploaded or created from scratch within Google Drive and the fact that everyone can access the file in one place means that it is much simpler to edit and update. This can be very useful for students and teachers who are collaborating on a project. This allows us to store the information in a single location where multiple administrators can edit and update it as necessary.

### Zoho

Another powerful tool for online collaboration is Zoho. It offers many online services and applications - free and commercial. Zoho Docs is online system for creating, storing and document management. It includes word processing, spreadsheet and presentations - applications. Additional tools which can be used for online collaboration in education are: Zoho Chat (communication), Zoho Mail (web based email), Project (planning and reporting activity for the project), Planner (for organizing daily tasks) and Zoho Wiki (for use with Wiki) and others. Working with these tools was possible for VFU-teachers during qualification courses. Teachers can implement tools from the Zoho platform in different programs. Zoho can import, upload and opens all common office file formats approved.

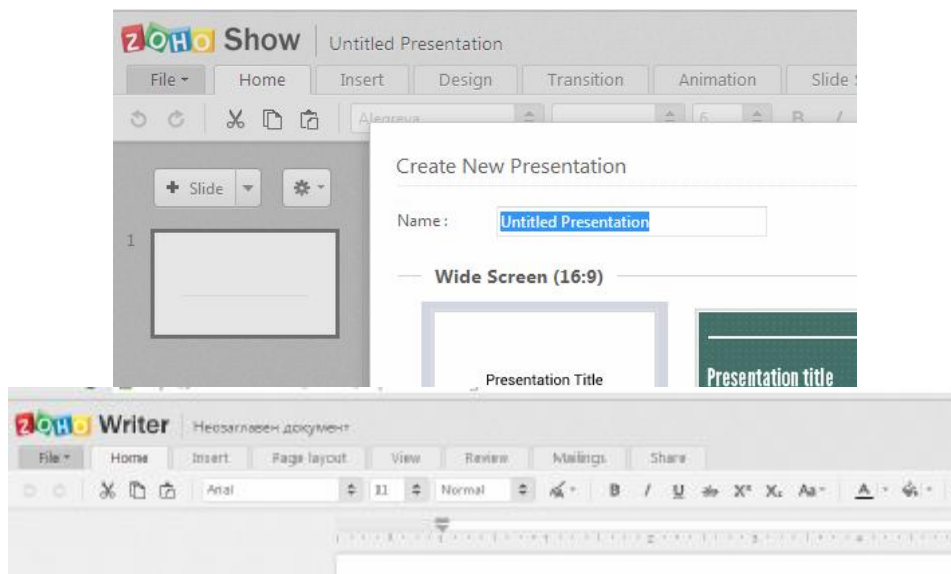


Figure 2. Creating documents and presentations with Zoho

Finished documents can be directly published to a blog or shared with colleagues / teammates, so to work simultaneously on the same document. Zoho Docs is an Online Collaboration tool, where is possible to create groups and invite friends, colleagues, etc., and share files with groups and collaborate in real-time. All group members have access to, and edit files online. Group collaboration is one of the main features in Zoho Docs which promotes teamwork, no matter where the team members are situated. Zoho Docs has a built-in Chat feature, which is useful to discuss with group members and share plans instantly. Restrict or permit specific rights to group members in the organization from anywhere. Manage your group members from a single spot.

Zoho Docs is integrated with Google Apps to make the document management more powerful, with more tools to help manage the documents better. With this integration the user gets Zoho Docs intuitive document management features and tools combined with Google's varied collaboration, storage and communication products and services.

With just a single sign-in and seamless integration, this will make it very easy for Google Apps customers to access all their Zoho Docs files and documents when logged into Google Apps.



Figure 3. Integration between Zoho and Google

### **Microsoft OneDrive (old SkyDrive)**

OneDrive is an online cloud storage service offered by Microsoft Company. It allows users to store, share and edit files, photos and more which are accessible anywhere in the world.

The desktop application OneDrive can synchronize files to any device automatically. The platform has ability to create office documents using Office Web Apps, simulates the interface of MS Office 2013, but with limited functionality. Accessing files from all devices and OS is possible - OneDrive app for the iPhone or iPad, Android or Windows Phone. OneDrive integrates with Microsoft's online version of Office, as did SkyDrive, but OneDrive adds true real-time collaboration. Previously, documents could be shared and edited among multiple people, but not simultaneously and seeing each other's edits. OneDrive is for true real-

time collaboration. Document versioning allows restoring earlier versions of the documents.

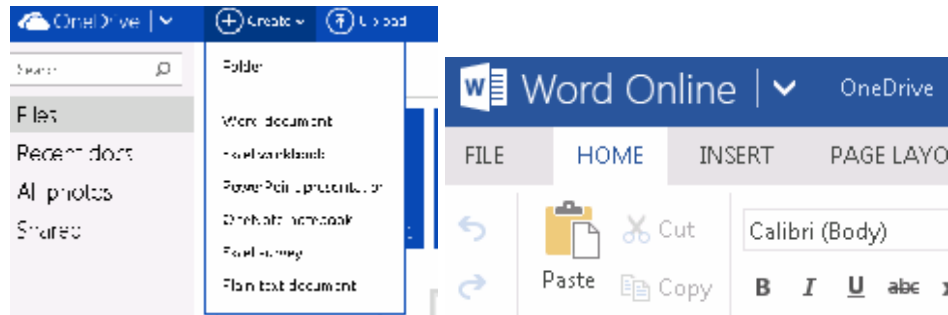


Figure 4. Creating documents with OneDrive

### Collaborating issues

Collaborating on documents is an excellent opportunity to work more effectively but sometimes it could lead to issues. When describing the positive results from online collaboration we shouldn't forget the problems which can occur. Knowing the problems we can optimize our work on document collaborating. Although the paper is about online collaboration in education, results from a business survey are relevant. **Version management** issues [3] have a greater hit on productivity than losing network connectivity or dealing with spam and include:

- Wasting time looking for the right file
- Manually merging edits from multiple contributors
- Working on the wrong or outdated document
- Sending the wrong document to a coworker or customer

Although 69% of knowledge workers<sup>1</sup> use a document management system or file sharing service, these respondents encountered document collaboration issues at the same rate, or even more frequently.

Other business and personal implications include poor business decisions from inaccurate information, loss of prestige, loss of clients, damage to personal reputation or working relationships, and even job loss.

### Conclusion

This paper highlights some platforms that could be used in educational process at Varna Free University as an online collaboration

<sup>1</sup> This report illustrates the effects of document collaboration issues quantified by a Harris Interactive online survey of 1 004 knowledge workers conducted in the United States and the United Kingdom during August 2012

tools for teachers and students. Online collaboration can significantly enhance teaching and learning. Based on faculty experiences, we also recommend starting small selecting one tool to pilot in one course. Knowing course requirements, platform possibilities and preventing problems with collaboration tools university professors can select the right online collaborative tool for their course and their students to make learning more attractive and engaging students.

### References

1. Brodahl Cornelia, Said Hadjerrouit, Nils Kristian Hansen. Collaborative Writing with Web 2.0 Technologies: Education Students' Perceptions. In: Journal of Information Technology Education: Innovations in Practice, Volume 10, 2011, p.73-103.
2. Ó Broin, Daire, Damien Raftery. Using Google Docs To Support Project-Based Learning. In: AISHE-J Volume 3, Number 1 (Spring 2011), p.00035.1-00035.11. <http://manchestergoogledocs.weebly.com/uploads/2/1/0/4/21046814/35-245-4-pb.pdf>
3. The Case for Better Document Collaboration. Findings from Harris Interactive® Knowledge Worker Survey, Commissioned by Perforce Software. <http://www.perforce.com/sites/all/themes/perforce/images/commons/perforce-document-collaboration-report.pdf>
4. Pollard Jim, Steve Burt. Online Collaborative Tools in Education. In: Clarity Innovations, December 2013 <http://www.k12blueprint.com/sites/default/files/Collaboration%20White%20Paper.pdf>
5. Bloom's Digital Taxonomy. Collaborative Editing using an Online Word Processor. <http://edorigami.wikispaces.com/file/view/collaborative+editing+using+online+Wp+rubric.pdf>
6. Ragupathi, K. Collaborative Learning using Google Docs&Maps: By Chris McMorran. In: Technology in Pedagogy, No 15, National University of Singapore, March 2013.

### Author:

**Antonina Ivanova Ivanova**, Varna Free University, Assist.Prof.  
[akrumova@abv.bg](mailto:akrumova@abv.bg)

## APPLICATION OF INFORMATION TECHNOLOGIES TO PRIMARY EDUCATION

Stefka Kovacheva

**Abstract:** *The article briefly describes the database, developed at the IMI-BAS for presentation of the UNESCO Worlds Cultural Heritage Sites in Bulgaria to pupils. This database will be used in an e-course project, currently under development, for integration of an additional e-learning content for the "Man and Society" course for 4th and "History and Civilization" course for 5th grade pupils. The project follows contemporary teaching strategy based on the formal structure of Understanding by Design approach, with emphasis on interaction between teacher, student(-s) and educational resources. In the process of knowledge acquisition students today are transformed from object of action into subject of action. This transformation requires mandatory education to be replaced by a new education environment in accord with tutees' needs and interests. Such an environment will develop their capabilities because tutees become active and productive participants in their own education.*

**Keywords:** *cultural heritage, curriculum and instruction design, database, e-course, education environment, understanding by design, UNESCO*

### Introduction

This article describes the *Database* presenting UNESCO World Cultural Heritage Sites in Bulgaria, which has been developed at IMI-BAS to provide an extended introduction and understanding of the national cultural and historical heritage for pupils. This Database has been planned and designed in order to integrate supplementary learning content into the "Man and Society" course for 4<sup>th</sup> and the "History and Civilization" course for 5<sup>th</sup> grade pupils.

Our project complies with the SER for the cultural and educational field of "Social sciences and civic education" for the "Man and Society" course for 4<sup>th</sup> and the "History and Civilization" course for 5<sup>th</sup> grade pupils [9, 10]. The project follows these standards: core (4<sup>th</sup> grade) learning content "cultural and historical heritage" – *Gives examples of significant cultural achievements of the Bulgarian people* and core (5<sup>th</sup> grade) learning content "national and world historical and cultural heritage" – *Identifies and describes typical monuments of the national and world culture.*

Besides the standards mentioned above the project also conforms with the standards of another core: the core (4<sup>th</sup> grade) learning content "knowledge sources" – *Differentiates between material and written knowledge sources; Composes short descriptions of objects and events;* and the core (5<sup>th</sup> grade) learning content "historical sources" – *Recognizes and distinguishes different types of historical sources;*

*Extracts information from texts and images; Composes short written texts with the help of accompanying questions; Formulates their own questions on historical problems and justifies answers; Becomes acquainted with the resources provided by ICT for searching and acquiring information.*

Furthermore an Internet based search shows that it is possible to find digital products online providing learning materials in the form of php-presentations [7], multimedia products (electronic games for children [12]), or educational websites [5, 11, 13]. These products provide electronic materials however they do not provide search options, addition of new information and resources, as well as direct learners participation to create and edit learning content.

The integration of new or supplementary learning content into the courses for pupils is the subject of an electronic education course – a project in the process of development. The project follows a contemporary teaching strategy, based on the Understanding by Design approach[4].

### **The learning process in the modern society**

The learning process in a technology based society is regarded as a permanently altered state, created as a result of experience and interaction with the environment. The search of information in the universe of the global network acquires a research profile as there is an ever increasing abundance of knowledge (information). The individual's ability to identify, comprehend and synthesize connections and models becomes a valuable skill. As mentioned by George Siemens [3], creator of the modern connectivism theory, "We derive our competence from forming connections".

**Connectivism** (networked learning) explains learning as a process of connecting specialized nodes or information sources in a network with the aim of expanding and developing thinking. The nodes can be communities, ideas, areas. Connecting unusual nodes promotes and enhances learning activity. The starting point of this theory is the individual (the learner). Personalized knowledge is enriched by the possibilities of the network to continuously update and transfer it which is the objective of every modern learning activity utilising technology.

**Education** is a symbiosis of teaching and learning, an altered state as a result of a permanent change in the human behaviour, which should

be a result of experience and interaction of the learner with the world (in our case – content or people).

**Teaching** is a process of influence over the learners, which provides for sufficient time to explore, consider and integrate the knowledge, presented by the tutor. It merges management and distribution of resources (designing the learning environment) in order to achieve the desired objectives and results.

**Learning** is a process, focused on connecting specialized information databases. The links between the different ideas or areas lead to the creation of new databases and supplementary connections, as well as the formation of the abilities necessary to distinguish between essential (relevant) and non-essential information.

Understanding is necessary in order to have active cognitive activity. Understanding the learning content is influenced by the method of presentation and the techniques of mental activity, mastered by the student.

**Comprehension** of the learning material is “a mediated analytic-synthetic process of disclosure through the mental consideration of essential properties of a particular cognitive object, distinction of its main components and their integration into a unified whole, whereupon the conditions and the “rules” of action become known” [6]. As a main component of the process of acquiring knowledge it guarantees establishment of connections between objects and the knowledge about these objects, i.e. comprehension is a person’s interaction with a particular object, resulting in “the re-creation (assembly) of a working model of the object in the subject’s mind” [6].

Knowledge, suggested solely by the teacher and the textbook is insufficient, because it does not provoke a thought process, it does not provide for creativity and their justified personal decisions. Modern education should acquire meaning in an interactive environment, which offers selectable forms of learning (“*whenever I am interested and as much as I need*”), diverse learning material and methods of interaction. The modern learning process utilises a large amount of information, which leads to perpetual dynamics in terms of the relation “cognition – practice – knowledge”; working in a network and collective intelligence – “*we too are creators of learning content*”.

## Understanding by Design

The ***Understanding by Design (UbD)*** approach is widely spread in the USA – hundreds of pedagogues utilize it as a tool for educational planning, focused on “teaching for understanding”. The emphasis of UbD is on “backward design” or more accurately put – “backwards planning”. This is a process which first identifies the outcomes in order to design curriculum units, performance assessments and classroom instruction.

**UbD** is a teachers’ handbook for designing and creating curriculum, based on standards, which “guides” the learners to understanding. The learners’ progress is monitored with the help of a program for testing and assessment of the acquired knowledge and experience.

Wiggins and McTighe define Understanding by Design as a “framework for designing curriculum units, performance assessments, and instruction that lead your students to deep understanding of the content you teach”. UbD suggests six facets of understanding, namely: students should explain, interpret, apply, have perspective, empathize and have self-knowledge about a given topic. UbD approach suggests two important premises: Backward design and Teaching for understanding.

### (a) Backward design

This is a three stage process where: (1) the teacher defines “what students will understand”, and lists essential questions, that will guide the learner to understanding – here the focus is on identifying “what students will know” and most importantly “what students will be able to do”; (2) the teacher sets tasks, whose performance will show what the learners have understood; (3) according to the cognitive changes that were demonstrated, the teacher suggests supplementary activities, which will lead the learner to the desired by the tutor results.

Essential questions and assignments predetermine the activities, which guarantee the comprehension and rationalization of the learning content.

### (b) Teaching for understanding

Teaching for understanding is another central premise of UbD, which should be evident in course design, teacher and learner attitudes and the classroom learning environment. There should be a coherent curriculum design and clear distinction between big ideas and essential questions. At the beginning of the curriculum unit (module) or the course the tutor should tell learners about big ideas and essential questions, about performance requirements and evaluative criteria. The learning



environment should offer incentives for all learners to come to understand the big ideas and answer the essential questions.

The Understanding by Design learning approach emphasizes on the learner's interaction with the learning content – a process of simultaneously active and interactive learning. In the process of acquiring knowledge pupils of today transform from objects of influence into subjects of action. This transformation suggests the substitution of learning out of obligation with a new learning environment, adapted to the needs and interests of the learners. Such an environment will develop their abilities, as it requires the learners to become active and productive participants in their own education by understanding “what” and “how” they learn and becoming aware of “why” and “how much” they know.

### **Our project**

E-learning is a process, which exists thanks to the substantial capabilities of the Information Technology (IT). The main advantage of E-learning is that it always ensures interaction between the learner, the teacher and the learning content, using the Internet and the IT as knowledge “suppliers”, the priority of self-control of the learner over the tutor's control.

Our aim, based on the academic approach “Understanding by design”, is to create an e-course, which will allow the learner to:

- utilize their own experience in order to complement the curriculum learning content;
- decide to edit/review the input data in order to create new knowledge output;
- establish the capacity to form links between data sources and between objects of knowledge, derived from this data.

The pupil aged 10-11 years is highly dependent on knowledge sources. For that reason with our project we suggest knowledge without the boundaries of the traditional methods of education – a modern form of learning and applying knowledge, based on the potential of the Internet.

The **big ideas** of our project, specially oriented towards the youngest pupils, are defined as follows:

- raising the educational level of young students;
- stimulation and extension of their interest towards history as a science;

- creating conditions for independent work and personal expression;
- informing them about UNESCO activities and the place of Bulgaria in world history and culture;
- building lasting notions of cultural assets and codes of conduct in order to preserve and protect them.

How to establish whether or not there is understanding?

**Evaluation criteria** of the results (recollection of the objects, understanding their value and symbolism, utilizing chronological and spatial cues, developing the ability to read documentary and artistic historical sources) are consistent with the standards set by SER [9, 10].

**The main purpose** is to expand our knowledge of culture and society in the context of the national and world heritage, by providing the opportunity to work individually with the included in the database objects, rich illustrative material and historical sources.

### **Architecture of the software environment**

The architecture of every software environment, designed to create and manage an e-course usually includes: modules, intended to ease the creation of learning content, links to repositories of learning material (databases) and users' interface. The base of the information model is the learning object (an independent part of the learning process, aiming at achieving a specific educational purpose). The repositories of learning objects are databases, containing learning materials (learning content in electronic format). The users of this medium are the tutors, the learners and their parents.

The creator of the program environment and its database is the so called administrator. The creator of the initial learning content is the tutor. The creator of the output learning content can be the tutor, the learners and their parents.

### **Main features of the software environment**

1) Administrator's activities – providing opportunities to control the learning process; providing connection with the repositories; operation of different types of users with corresponding access rights; provision of means of search and communication;

2) Tutor's activities – design the learning process; creation of learning modules, related to the database objects and assessment assignments – questionnaires, tests, quizzes;

3) Learner's activities – comprehend and complete the assessment assignments while using the training modules to acquire data; create output learning content alone or in a group based on the set criteria;

4) Parents' activities – monitor their children's individual work and progress;

### **Technology platform**

The Web 2.0 platform has been selected for the implementation of the project – an e-course to aid the pupils in the fourth and fifth grade (in development). Our idea is to present the new or supplementary learning content as a set of hyper-documents, combined with multimedia components, delivered as learning materials via Internet.

### **Database presenting UNESCO World Cultural Heritage Sites in Bulgaria**

Why was the Database developed?

**Because** we presume that at this age (when the children have less experience and great enthusiasm to seek expression outside the textbook) it is appropriate to gain intimate knowledge of the values that provoke a sense of pride in our Homeland;

**Because** the objects hold an exceptional cultural and historical significance to the common heritage of mankind, thus they should be promoted and be preserved for future generations; all this needs to be known and remembered!

**Because** throughout the duration of the lesson IT provides the means necessary for each child to “see” the objects they are being taught about probably for the first time.

### **World Heritage Sites in Bulgaria protected by UNESCO**

UNESCO World Cultural Heritage Sites in Bulgaria include seven architectural and historical sites. The first four properties meeting the criteria set in [14] were inscribed in the list with sites under UNESCO protection in 1979: № 42 Boyana Church, № 43 Madara Rider, № 44 Thracian Tomb of Kazanlak and № 45 Rock-Hewn Churches of Ivanovo. The next two sites were inscribed in 1983 - № 46 Rila Monastery and №

47 Ancient City of Nessebar. The final site – Thracian Tomb of Sveshtari – was inscribed in 1985.

1. Boyana Church

Inscribed in the UNESCO list [15], because it meets two criteria, in accordance with [14]: 2 and 3. This is one of the few completely preserved medieval monuments with unique medieval frescoes from 1259, also famous for the first image of St. John of Rila.

2. Madara Rider

The only rock relief in Europe dating back to the early Middle Ages, carved VIII – IX century. It meets two criteria in accordance with [14]: 1 and 3.

3. Thracian Tomb of Kazanlak

It meets three criteria in accordance with [14]: 1, 3 and 4. It is a masterpiece of the Thracian culture dated IV – III century BC. It is the oldest tomb in Europe with the best preserved frescoes – samples of ancient art from the Early Hellenistic Age.

4. Rock-Hewn Churches of Ivanovo

The Rock-Hewn Churches of Ivanovo shape a monastery, sample of the medieval Bulgarian art, created XI – XIV century. This is a site that meets three criteria according to [14]: 7, 8 and 9. The monastery is a complex composed of small rock-hewn churches, chapels and cells, the height of which ranges between six and ten meters. Well preserved unique frescoes from the XIII and XIV century are some of the most significant examples of monumental painting on the Balkans.

5. Rila Monastery

The largest monastery complex in Bulgaria and the second largest on the Balkan Peninsula (the first being Aton) was founded in the X century by John of Rila. A monument of the Bulgarian medieval and national revival architecture and painting, it meets one criterion in accordance with [14]: 6. It is one of the most important cultural, educational and literary centres in Bulgarian history.

6. Ancient City of Nessebar

It is inscribed in the list because it meets two criteria: 3 and 4. One of the oldest towns in Europe – it was founded more than 3000 years ago as a fortified Thracian settlement named Mesambria. An architectural ensemble of the ancient wall from the V – IV century BC, Early Byzantine churches from the X – XIV century, over 60 National Revival Period houses from the XVIII century. The Nessebar churches have their own picturesque architectural style and rich external decoration.

7. Thracian Tomb of Sveshtari

Thracian-Hellenistic royal tomb from the III century BC; it blends culture, art and architecture. It meets two criteria in accordance with [14]: 1 and 3.

### **Structure and attributes of the described sites**

According to [14] the observation of the UNESCO sites lays emphasis on certain characteristics (such as bearers of values, shape, material, location, etc.). The elements comprising the Database (i.e. the object of the database) are the sites in Bulgaria, placed under the protection of UNESCO.

There are five types of objects: *city* (Ancient City of Nessebar); *monastery* (Rila Monastery); *church* (Boyana Church, Rock-Hewn Churches of Ivanovo); *tomb* (Thracian Tomb of Kazanlak, Thracian Tomb of Sveshtari); *relief* (Madara Rider).

The Database is presented in a tree structure. All objects in the Database are complex objects. In order to describe an object of the Database, we use metadata – a great number of attributes, set on several levels, which have specific values. The values of the attributes are texts and images, containing learning content.

The attributes used to describe an object in the Database are divided into three groups.

The first group is comprises the attributes that correspond 100% to the characteristics described in [14].

The second group includes the attributes, which hold new (supplementary) knowledge, which is not referred to in [14], however it will certainly wake the interest of this age group (history of the monument, depiction in literature, artefacts, found at the site, etc.).

The third group contains attributes, which are specific to certain object types (frescoes – churches and monasteries, environment – Madara Rider and Rock-Hewn Churches of Ivanovo).

The Figure 1 shows the tree structure of database objects. The Figure 2 shows a screenshot *Category of “history”: marked sample button “history” in the subcategory “town”* – Ancient City of Nessebar.

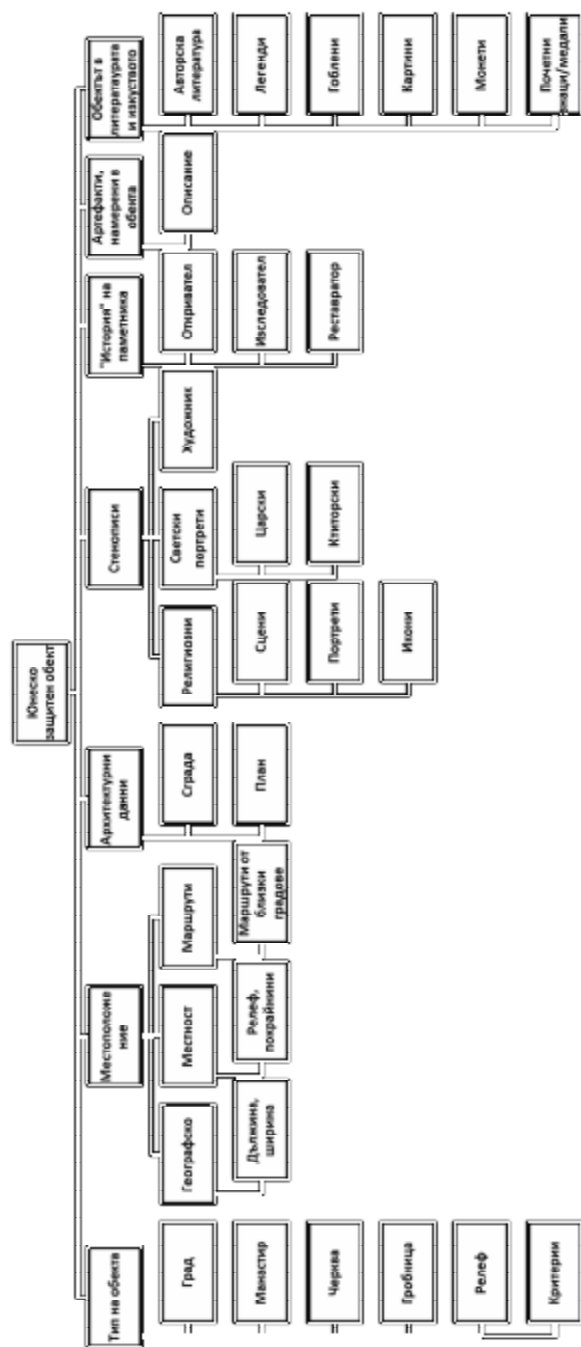


Figure 1. Tree structure of database objects



Figure 2. Category of “history”: marked sample button “history” in the subcategory “town” – Ancient City of Nessebar

## **Conclusion and future development**

This article discusses the principles of UbD, intended to provide a more effective and engaging learning process. The UbD framework utilises an approach which illustrates the continuous improvement of both pupils' achievements and teachers' professional skills. The results of modeling - pupil achievements - are the feedback for eventually taking action related to corrections in the proposed by us model in terms of quality and level of learning process. Thus can be achieved expected from us higher level of understanding information from learning resources.

The Database presenting UNESCO World Cultural Heritage Sites in Bulgaria allows for expansion by including new sites from the Bulgarian cultural and historical heritage (e.g. The Red Church near Perushtitsa, Architectural reserve Old Plovdiv, etc.). This extended database can be used in the creation of a website, dedicated to the youngest students. The children will be able to present regional cultural monuments and exchange information, concerning the perception and understanding of Bulgaria's cultural values. This way the children become co-authors of content, generate experience, and realizing the significance of the world heritage under UNESCO's protection will motivate them to preserve these monuments.

The international organization UNESCO emphasizes the need to improve the quality of education through the introduction of better practices for teachers and the mobilization of more resources.

UNESCO's mission is to provide universal access to information and information technology in order to promote learning and cooperation, conducive to the sharing of common values, to assist understanding, interpreting and preserving of the cultural environment.

## **References**

1. Cachia, R., Ferrari, A., Ala-Mutka, K., and Punie, Y. Creative Learning and Innovative. Teaching Final Report on the Study on Creativity and Innovation in Education in the EU Member States. Institute for Prospective Technological Studies, Luxembourg: Publications Office of the European Union, 2010
2. Moore, M. (Working Group Coordinator). Information and Communication Technologies in Distance Education. Specialized Training Course, UNESCO Institute for Information Technologies in Education, 2002
3. Siemens, G. Connectivism: A Learning Theory for the Digital Age. Elearnspace, December, 12, 2004. (<http://www.connectivism.ca/> update April 5, 2005)
4. Wiggins, G., J. McTighe. Understanding by Design framework. Expanded 2<sup>nd</sup> ed. 2012. ([http://www.ascd.org/ASCD/pdf/siteASCD/publications/UbD\\_WhitePaper0312.pdf](http://www.ascd.org/ASCD/pdf/siteASCD/publications/UbD_WhitePaper0312.pdf))



5. Българско училище в чужбина. Блогът на Илиана Т. Илиева-Дъбова за българска история, българска география, етнография, изкуство. Образователен блог. (<http://ilieva-dabova.blogspot.com/> )
6. Десев, Л. Речник по психология, Булгарика, С., 2010
7. Дунчева, Петя. Електронен учебник „Човекът и обществото“ за 4 клас
8. Мутафов, С., Иванов, И. Педагогическа хигиена. Шумен, 1996.
9. SER: Държавни образователни изисквания за учебно съдържание по “Човекът и обществото”: <http://www.paideiafoundation.org/mce/file/DOI%20chovek&obshtestvo%20I-IV%20klas.pdf>
10. SER: Държавни образователни изисквания за учебно съдържание по “История и цивилизация”: <http://www.paideiafoundation.org/mce/file/DOI%20istoria%20V-VIII%20klas.pdf>
11. Електронни учебници по “История и цивилизация - виртуален свят за децата със специални образователни потребности” (<http://www.history-sop.com/>, <http://www.youtube.com/watch?v=h4f9uVmAqsY>)
12. Мултимедиен продукт „Открий съкровището“ (<http://otkrivam.com/> )
13. Уча се. Образователен сайт. (<http://ucha.se/videos/istoriya/5-klas/> )
14. UNESCO World Heritage Centre - The Criteria for Selection: <http://whc.unesco.org/en/criteria/>
15. <http://unesco-objects.hit.bg/>

**Author:**

**Stefka Kovacheva**, MSci  
PhD Program of the IMI–BAS  
Mobile: +359 888466063  
E-mail: [kovachevastefka@gmail.com](mailto:kovachevastefka@gmail.com)

## GRADING SYSTEMS IN TEACHING OF PROGRAMMING

Krassimir Manev

**Abstract.** *Competitions in programming appeared in the 70s of the past century for attracting talented young people to the profession of programmer. Soon after the start of organizing programming contests, grading systems (GS) become an inevitable part of the process of evaluation of contestants' results. Recently, different applications of the grading systems were announced that were going beyond the initial scope of these systems. One of the possibilities is to use GS in the teaching of programming. The paper presents the main objectives and concepts in the domain, examples of GS developed by the authors and the work that is necessary to do for adapting GS to educational process.*

**Keywords.** *teaching programming, correctness of programs, grading systems.*

### I. INTRODUCTION

The usage of computing systems (computers and corresponding software) in modern education is inevitable. In education, computers are intensively used nowadays to prepare and distribute teaching materials, to search information necessary for elaborating of assignments, for examination, for maintaining the contact among the teacher and students, and other educational activities. The complex of all these activities, denoted by e-learning, is a topic of very active research and development.

It is a paradox that, in contrast with the other domains, computing systems are **less used to help educational process in the domain of Informatics** (we will denote in such general way the different sub-domains connected with creation and applying of computing systems – Computer engineering, Computer science, Information systems and Software engineering). May be because computing systems are the central objects of the teaching in the domain, their possibilities to help teachers and students in educational process are rather neglected – the specialists prefer to develop applications that help education in other disciplines than in their own. This situation could be dramatically changed.

From all courses of Informatics the most appropriate for applying software systems in educational process are courses teaching programming and algorithms. The extracurricular activity that is closest to teaching of programming and algorithms is the *competitive programming*. Programming contests has more than 30 years of history. Programming contests are usually organized by very experienced professionals. That is why a huge amount of resources, including software tools, is created to help organization of programming contests. All these tools could be adapted for teaching programming in universities as well as in secondary schools.

This paper consider the possibilities to use one of the instrumental tools for organizing competitions in programming – so called grading systems – for teaching programming courses. Section 2 presents the programming contests and the possibilities of the grading systems. Three examples of such systems developed and used by authors are given. In Section 3 some challenges that education of programmers put in front of the teachers are outlined. For each challenge it is described how the grading systems could help. Section 4 concludes about the efforts that are necessary to adapt existing grading systems for usage in education.

## **II. PROGRAMMING CONTESTS AND GRADING SYSTEMS**

Competitive programming is the activity that is closest to teaching of programming in secondary schools and universities. The essence of organizing programming contests is really preparation of programmers with very high competence, going far away beyond schools' and universities' curricula. The positive role of the scientific contests, and especially of the contests in mathematical disciplines, was outlined long time ago [1]. Two of the main objectives for such estimation are that:

- it is more attractive for young people when the education is organized in parallel with and for the purposes of the participating in competitions;
- many of the challenges that the children will meet in their adult life have a form of a competition, so it is necessary especially to teach students to compete. It is better to start the process as early as possible.

### **1. Programming contests**

In 1977 the Association of Computing Machinery starts its Collegiate Programming Contest for students in the American Universities. Very soon the contest was extended to an international event. Nowadays the International Collegiate Programming Contest of ACM attracts students from more than 1700 universities and more than 80 countries all over the world.

Bulgaria is among the pioneers of the programming contests for school students. First programming contests for school students in Bulgaria were organized in the early 80-th of the past century. In May 1989, Bulgaria organized and hosted the First International Olympiad in Informatics (IOI) for school students [2]. Soon, some regional Olympiads in programming appeared too – Balkan OI, Central-European OI, Baltic OI, etc. In 2007 the first Olympiad for school students aged less than 15.5 years took place in Belgrad, Serbia – the Junior Balkan Olympiad in Informatics.

Recently, many events different from Olympiads are organized also. Understanding well the role of the contests, some professional companies invest money and maintain their own teams for preparing tasks, for organizing contests and evaluating them. Examples of such contests are Top Coder and Google Code Jam. Similar activities are the various web sites that propose a continuous on-line training process with competitive elements. Very popular training sites nowadays are USACO Training Gateway, UVa Online Judge, Timus Online Judge and Russian Open Cup (*Открытый кубок по программированию*) – readers that are interested in will find easily the sites of these contests in Internet.

During a programming contest each contestant has to solve one or more *tasks*. The statement of the task usually contains an object (or set of objects) with given properties called *input*. The contestant has to write a computer program, called *solution* of the tasks which is able, for given (but unknown by contestant) input, to find a new object (or objects), being in prescribed relations with the input object(s) called *output*.

The main feature of the solution that is usually evaluated is its *correctness*. Obviously, the size of the input space could be enormously large and it is not possible to check the program with all possible inputs. That is why the author of the task has to choose for testing the correctness of the program some subset of inputs called *test cases*. Submitted programs are checked on test cases and the obtained outputs are compared with the supplied by the author outputs. The other feature of the solution is its *speed (time complexity)*. The speed of the programs is controlled by a specific for each task *time limit* – an interval of time inside which the execution of the program has to finish.

A problem of such kind of testing is that some tasks allow more than one correct output for some inputs. In such cases the correctness of the output has to be checked by a special program called *checker*.

## 2. Grading systems

From the mentioned above it is clear that the checking of correctness and effectiveness of computer programs is a complex set of activities. Some of these activities could be performed only with usage of appropriate software. Other could be performed manually but this is boring, time consuming and risky. In the end of 80's and beginning of 90's years of the past century specific software systems were introduced to help evaluation of programming contests called *grading systems* (GS). Nowadays GS are inevitable part of the programming contests and training sites.

Grading system incorporates all elements that are necessary for organizing the contest – tasks' statements, time limits, input test cases,

expected outputs, checkers, contestants' programs, results of evaluation, etc. GS receive submitted by contestants programs, compile them, run them on the input test cases, check the result and assign evaluation marks in accordance with defined in the system evaluation policy. Depending of the policy GS could inform the contestant for her/his results or part of results. GS could archive different versions of contestant's programs, to execute program on prepared by contestant test cases, etc.

Modern GS do not share common history. In fact most are developed as silos from their beginning. The reason is that a simple GS is relatively easy to build, if not easier than understanding an already existing one. Moreover some of the best known GS are with *closed code* – their code is not publicly available and no guidance is provided for their setup. Thus most people that need to grade some programs outside of the software engineering process end up with creating their own GS. The authors are involved for many years in organizing programming contest, as well as in design and development of different grading systems, namely:

**SMOC** (System for Managing Online Contests) is the GS used in all national programming competitions for high school students in Bulgaria. Originally it was a modification of the grading system of IOI'2002 (Republic of Korea) for the purposes of Balkan OI'2004 (Bulgaria). It has since grown and the original system is virtually completely rewritten. Since then it was also used in Balkan OI'2006 (Cyprus), Balkan OI'2008 (FYRO Macedonia), JBOI'2008 (Bulgaria) and IOI'2009 (Bulgaria). Aside from the on-site contests it is also used to handle simultaneous online competitions.

The main advantage of SMOC is the safety evaluation process and precise time measurement developed in cooperation with the Technical Committee of IOI [3],[4]. The modern version of SMOC is able to support several types of tasks. However it is still closely coupled with IOI-like contests – each test case is checked separately. SMOC has not the functionality to be an archive of tasks and to be used for training. Tasks evaluated with SMOC during national contest are collected and archived in a web site not linked with the usage of the system.

**spojo** is a GS designed to work continuously [5]. It is able to host several contests at a time, and allows online and offline submits. It was used for organizing contests in Sofia University, as well as for checking exam and homework assignments in several (algorithmic) courses in the university. Evaluation provided is in ICPC style – each submission is accepted if passed successful all test cases. The system has an archive of used tasks. The tasks are archived in groups in the way they were proposed for training, homework or examination contests. The system

has an installer for Debian GNU/Linux based systems and exporter for plagiarism checkers.

Because of the very simple design, spoj0 uses only tools of the operating system to provide secure execution and has not so precise time measuring. System administration is performed by command line scripts on the hosting machine and need some experience.

**Maycamp Arena** is a site that is dedicated especially for on-line training and organizing of on-line programming contests [6]. The GS of the site is a cloning of spoj0, maintaining an evaluation in IOI style. Archived tasks are classified by the domain of the used algorithm so each participant could choose task from the selected topics. Very important feature of Maycamp Arena is the collection of statistic data about performance of contestants and hardness of the tasks.

The site has a small history and is still under development and amelioration.

More detailed description of the existing programming contests and used GS could be found in [7].

### III. CHALLENGES OF TEACHING PROGRAMMING

The teaching of programmers is very similar to the education of artists – for example, piano players. Teaching of programmers starts with learning of necessary instruments – one computing system and one programming language as well as the teaching of piano players start with learning of the **instrument** (the piano) and the musical **notation**. At this stage, before the start of creation of own programs, it is very helpful for the programming beginners to look at some program code **samples** written by experienced programmers and to try to understand them, as well as it is helpful for young piano players to listen some musical pieces interpreted by experienced artists.

Then the education of piano players continues with playing a set of **etudes** – small musical pieces especially written for the purposes of education. Each etude contains some technical element, or composition of such elements, which are important for the development of the player. In a similar way, education of the programmers continues with solving of many programming etudes – small programming tasks. For solving each programming etude the student has to demonstrate good knowledge of some programming technique, ability to implement some abstract data type in data structures and to apply some algorithm.

The last stage of education of the piano players is preparing to play large musical forms – concerts or big solo parts of symphonies. Programmers also have to pass such stage of education in order to be prepared to implement relatively large **software systems**. On this stage

they learn how to decompose programming projects in reasonable parts and to compose the final software product from implemented parts, how to control the quality of the implementation, how to work in teams, and so on. All this subjects taught in the discipline Software technologies and will not be discussed here.

GS could be very useful for teaching of programming because it is a part of training of contestants. From the other side, teaching of programming includes evaluation of the results of the students and this is one of the most challenging parts of the educational process.

In many schools and universities the teachers still are checking manually program fragments written by the students, as homework or exam works, on paper. This is an activity which could be totally assigned to GS. The most natural usage of GS in education is during the heavy university courses of Algorithms. We are really using the mentioned above system spoj0 for many years in Faculty of Mathematics and Computer science of Sofia University for evaluating home works and exams of the course of Algorithms.

It is amazing that GS could be used for evaluation in some very early stages of education – when the students have not the knowledge and experience to write complete program [8]. In such stage they are asked to write program fragments – functions, procedures, classes of simple objects, etc. – and the checking of the assignments could be done with GS if the statements of the tasks are prepared as interactive. The missing part, that students are not able to write have to be prepared as author module and to be integrated with the solution during the evaluation process. The authors of [9] are going further – they are using a GS not only for teaching classical courses of procedural programming but for courses of Logic Programming too.

Finally, it is possible to use the experience of the competitive training sites – GS together with a large archive of tasks could be published in Internet. All homework and exams could be made on-line by the students. And more, after expiring of the deadline of the homework or the exam, the students, that did not succeed to solve some of the tasks, could continue to work on them till obtaining a correct solution. In such way using the GS for education in programming make it more flexible, intensive and efficient.

Below we will consider the main challenges for the teacher of programming in order to identify how GS could be used in educational process:

### **1. Necessity of large amount of samples and etudes**

Necessity of large amount of programming samples and etudes is one of the main challenges of teaching programming. After long period of working each teacher could succeed to collect own set of samples and etudes but in the beginning of the career it is a real difficulty. Something more – programming is very dynamic discipline and the actual platforms and approaches sometime change dramatically. After such changes the teacher has to reconsider in depth her/his set of samples and etudes.

A huge amount of samples and etudes could be taken from archives of existing GS. For the purpose the archived tasks and solutions have to be appropriately classified. Tasks have to be classified by their domain and hardness, and solutions – a by their effectiveness. Such classifications could be based on the statistics collected during the exploitation of GS. In the above mentioned systems we have some elements of the necessary functionality but they are not well developed. Unfortunately, classification of the tasks could not be made automatically and will need serious efforts.

### **2. Necessity of checking large amount of students' assignments**

Checking a large amount of programming assignments is the greatest challenge for the teacher. Just reading the program code and understanding the idea of the program is difficult. Finding the mistakes is practically impossible. And here is the greatest advantage of GS. They could check thousand assignments, to evaluate them and to produce the corresponding report inside a minute.

### **3. Necessity of checking different kinds of students' assignments**

Depending of the stage of the educational process different kind of evaluation has to be applied. For examination it is appropriate evaluation in ICPC style – the assignment is accepted if solution pass all of the tests. The IOI style, for example, is more appropriate for home works and midterm quizzes – in these cases the system could assign points for tests that program passed successfully. For exercises in class it will be helpful for the student to receive from the system not only the number of problematic test cases but the inputs for these tests also in order to debug the program. In such case, if necessary, the system could provide the correct output too.



All these kind of checking are provided by some of mentioned above system. Something more – the corresponding kind of evaluation could be applied consecutively, decreasing the strictness, in order to estimate precisely the level of knowledge and ability of the student. In cases when the student is not able to solve the task and if the archive of classified tasks is well developed and maintained then GS could recommend some more easy tasks to be solved as a preparation before the student to return to the difficult task.

#### **4. Necessity of applying more flexible and adequate forms of evaluation**

Evaluation of the students' results manually is very rough – usually the teacher is able to assign and check 2-3 home works and 1-2 quizzes. In such way the grade of the student will depend too much of the choice of the tasks for the assignments and will not be adequate enough to the knowledge and the abilities of the student. Using GS the teacher could check practically each of the learned topics. And more, based on the statistics for hardness of different topics, collected in the system, she/he will be able to formulate grading criteria which to lead to a grading mark much more adequate than the obtained without using GS.

One other problem of the evaluation that could be solved with GS is the plagiarism, which despite the efforts of the teachers still exists. As we mentioned above it is enough to embed in the system some of the existing system for discovering of plagiarism. Some of these systems are sophisticated enough and able to discover much more than the trivial Copy + Paste plagiarism.

#### **IV. COMMENTS AND CONCLUSIONS**

Modern grading systems, designed to support organizing of programming contests, could successfully and fruitfully be used to help teaching of programming. As shown provided above analysis the following steps will be necessary in order to apply GS actively in the teaching process:

1. To develop a *repository* for archiving tasks and necessary statistics. One possible organization of the repository of the tasks is given in [10] but it is not enough for our purposes. The functionality for collecting statistics of UVa training site, for example [11], is very well designed and implemented and could be used as a sample for further implementations;

2. To develop a module which implement the notion *contests* (a generalization of the different possible teaching activities – exercises,

home works, individual trainings and examinations). The module will provide to teachers and students the possibility to define contests depending of their needs and will be an integrator of the complex of systems distributing the evaluation of different task of each contest among the existing GS;

3. To develop appropriate *user interfaces* for teachers and students that will give them possibilities to define necessary contests without necessity to now inner organization of the GS and the repository.

Based of this conclusion two system architectures was designed and are under development but their discussion is going beyond the scope of this paper.

### References

- [1] Verhoeff, T. The Role of Competitions in Education. Presented at *Future World: Educating for 21<sup>st</sup> Century*. A conference and exhibition at IOI'1997.
- [2] Kenderov, P., Maneva, N., ed. *International Olympiad in Informatics*, Sofia, 1989.
- [3] Mareš, M. Perspectives on grading systems. *Olympiads in Informatics*, v. 1, 2007, 124-130.  
[http://www.mii.lt/olympiads\\_in\\_informatics/pdf/INFOL003.pdf](http://www.mii.lt/olympiads_in_informatics/pdf/INFOL003.pdf)
- [4] Tochev, T., Bogdanov, Ts. Validating the Security and Stability of the Grader for a Programming Contest System, *Olympiads in Informatics*, vol. 4, 2010, 113-119.  
[http://www.mii.lt/olympiads\\_in\\_informatics/pdf/INFOL058.pdf](http://www.mii.lt/olympiads_in_informatics/pdf/INFOL058.pdf)
- [5] Средков, М. *Система за организиране на състезания spoj0*. Diploma thesis for obtaining MSs degree, Faculty of Math. And Comp. Science, Sofia University, 2006.
- [6] Михов, В. Проектът мейкамп арена – използване на състезателни системи в обучението по програмиране, *Mathematics and Education in Mathematics*, Proc. of 40-th Spring Conference of UBM, Borovetz, 2011, 438-443.  
[http://www.math.bas.bg/smb/2011\\_PK/tom/pdf/438-443.pdf](http://www.math.bas.bg/smb/2011_PK/tom/pdf/438-443.pdf)
- [7] Manev, Kr., Sredkov, M., Bogdanov, Ts. Grading Systems for Competitions in Programming. *Mathematics and Education in Mathematics*, Proc. of 38-th Spring Conference of UBM, Borovetz, 2009, 103-116.  
[http://www.math.bas.bg/smb/2009\\_PK/tom\\_2009/pdf/103-116.pdf](http://www.math.bas.bg/smb/2009_PK/tom_2009/pdf/103-116.pdf)
- [8] Ribeiro, P., Guerreiro, P. Early Introduction of Competitive Programming. *Olympiads in Informatics*, v. 2, 2008, 149-162.  
[http://www.mii.lt/olympiads\\_in\\_informatics/pdf/INFOL020.pdf](http://www.mii.lt/olympiads_in_informatics/pdf/INFOL020.pdf)
- [9] Ribeiro, P., Simões, H. and Ferreira, M. (2009). Teaching Artificial Intelligence and Logic Programming in a Competitive Environment. *Informatics in Education*, v. 8(1), 85-100.  
[http://www.mii.lt/informatics\\_in\\_education/pdf/INFE143.pdf](http://www.mii.lt/informatics_in_education/pdf/INFE143.pdf)

[10] Verhoeff, T. Programming Task Packages: Peach Exchange Format. *Olympiads in Informatics*, v. **2**, 2008, 192-207.

<http://olympiads.win.tue.nl/oi/oi97/ffutwld/competit.html>

[11] Revilla, M. A., Manzoor, S. and Liu, R., Competitive learning in informatics: the UVa on-line judge experience. *Olympiads in Informatics*, v. **2**, 2008, 131-148.

[http://www.mii.lt/olympiads\\_in\\_informatics/pdf/INFOL035.pdf](http://www.mii.lt/olympiads_in_informatics/pdf/INFOL035.pdf)

**Author:**

**Krassimir Nedelchev Manev**, Assoc. Professor, Ph.D.  
Department of Informatics, New Bulgarian University  
(+359) 887 253 211, [krmanev@nbu.bg](mailto:krmanev@nbu.bg)

## MAPLE AND MATLAB IN TEACHING DEFINITE INTEGRAL

Zlatka Mateva, Nedka Pulova, Violeta Bojikova

**Abstract:** *This article presents practical experience on the application of the Computer Algebra Systems (CAS) Maple and MATLAB in teaching process on the topic “Definite integral”, included in all engineering curriculums at Technical University of Varna. In addition to visualizations, which facilitate the understanding of the mathematical definition of the definite integral, the article also offers a set of problems and their solutions as well as comments on these solutions from a didactic perspective. Emphasis is on the possible positive and negative effects of CAS’s implementation in the process of teaching definite integral and its applications.*

**Keywords:** CAS, definite integral, teaching, education

### Introduction

The emergence of computers in the mid 20<sup>th</sup> Century has begun to transform the teaching and learning. Computer Algebra Systems (CAS) are getting more and more incorporated in university-level mathematics education. However, little is known about the extent of CAS use and the factors influencing its integration into university curricula. Mathematicians frequently talk about CAS and CAS-assisted teaching in relation to their mathematical views and principles. We believe that CAS offers a number of new ways for teaching mathematics. CAS can be used to encourage discovery and experimentation in classroom. We know that the use of CAS in lab-classes can help to make the mathematic objects understandable and helpful. The visualization feature of CAS can be effectively employed in teaching, to generate conjectures.

For more than ten years the computer algebra systems Maple, Mathcad and MATLAB are used in education of mathematics in the form of lab-classes in the Technical University of Varna. Our colleagues use CAS-based lab to

- visualize graphics and project images;
- assist experimentation, exploration, and research in classes;
- offer realistic, complex, or real- world problems for students;
- enable them to devote more time for conceptual problems;
- motivate students in classes;
- encourage group work in classes;
- prepare and offer homework assignments;
- check student assignments and worksheets.

Here we present a small part of our experience in teaching definite integrals in lab-classes in Technical University of Varna.

### Riemann sums

In the corresponding lecture the definite integral of  $f(x)$  is defined as the limit of Riemann sums

$$\int_a^b f(x)dx = \lim_{\substack{n \rightarrow \infty \\ \Delta x_k \rightarrow 0}} \sum_{k=1}^n f(x_k^*) \Delta x_k.$$

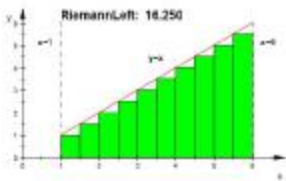


To help students more readily adopt this definition we can firstly estimate one simple definite integral  $\int_a^b f(x)dx$  by Riemann sums

$R_n^* = \sum_{k=1}^n f(x_k^*) \Delta x$ , where  $n$  is the number of subintervals and for example

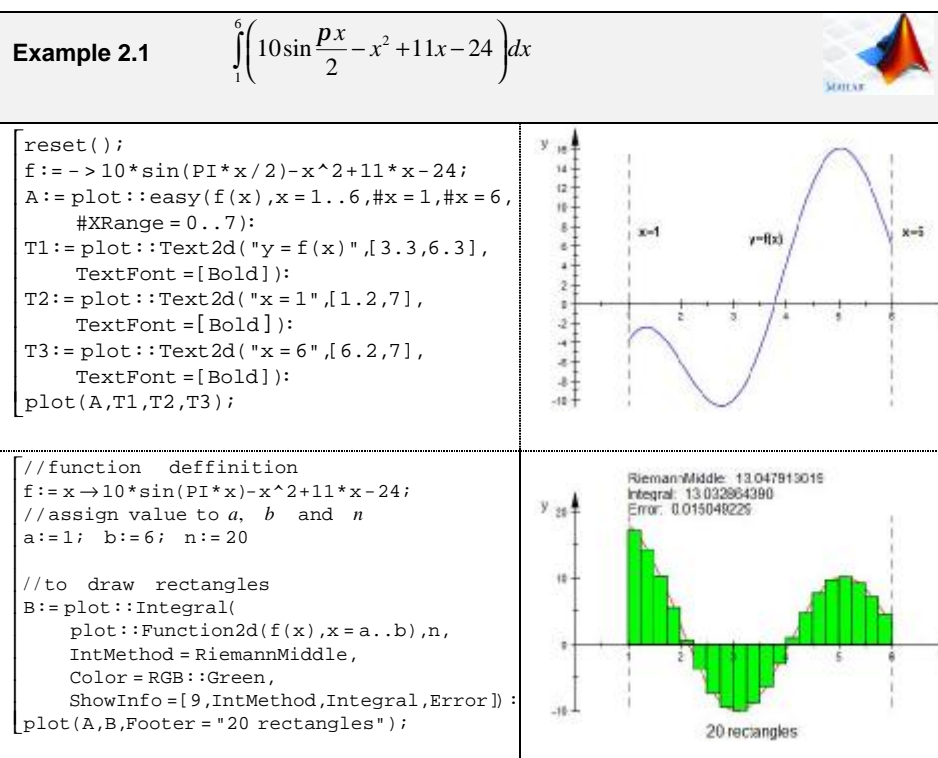
equals to 10, 20, 13 and 60,  $\Delta x = \frac{b-a}{n}$  is the length of each subinterval

and  $x_k^*, k=1,2,\dots,n$  are selected points in each subinterval. Some popular choices of  $x_k^*$  are: the left end point ( $x_k^* = a + (k-1)\Delta x$ ), the right end point ( $x_k^* = a + k\Delta x$ ) and the midpoint ( $x_k^* = a - \frac{\Delta x}{2} + k\Delta x$ ) of each subinterval.

**Example 1.** Calculate the left, right and middle Riemann sums to the integral  $\int_1^6 x dx$ .

<p>a) <math>n = 10, \Delta x = \frac{6-1}{10} = \frac{1}{2}</math></p> <p>and choose <math>x_k^*</math> to be left endpoints</p> <p><math>x_k^* = 1 + (k-1)\frac{1}{2} = \frac{k+1}{2}</math></p> <p><math>R_{10}^{left} = \sum_{k=1}^{10} \left[ \frac{k+1}{2} \cdot \frac{1}{2} \right] = 16,25</math></p>  <p><math>R_{20}^{mid} = 17,5</math></p>	<p>b) <math>n = 10, \Delta x = \frac{6-1}{10} = \frac{1}{2}</math></p> <p>and choose <math>x_k^*</math> to be right endpoints</p> <p><math>x_k^* = 1 + k\frac{6-1}{10} = \frac{k+2}{2}</math></p> <p><math>R_{10}^{right} = \sum_{k=1}^{10} \left[ \frac{k+2}{2} \cdot \frac{1}{2} \right] = 18,75</math></p>  <p><math>R_{30}^{mid} = 17,5</math></p>	<p>c) <math>n = 10, \Delta x = \frac{6-1}{10} = \frac{1}{2}</math></p> <p>and choose <math>x_k^*</math> to be middle points</p> <p><math>x_k^* = \frac{3}{4} + k\frac{1}{2} = \frac{2k+3}{4}</math></p> <p><math>R_{10}^{middle} = \sum_{k=1}^{10} \frac{2k+3}{4} \cdot \frac{1}{2} = 17,5</math></p>  <p><math>R_{60}^{mid} = 17,5</math></p>
--	---	--

£ That problem is not difficult and Students can easily calculate the Riemann sums above. Of course, in difficult cases, this leads to messy computations, as the closed form of Riemann sum is, in general, very hard to find. Maple and MATLAB/MuPAD possess commands designed especially for calculation and graphical presentation of Riemann sums (command **plot::Integral()** in MuPAD; commands **RiemannSum()** and **ApproximateIntTutor()** in the subpackage **Student[Calculus1]**, and some operators in the old package **student** as **leftbox()**, **leftsum()**, **rightbox()**, **rightsum()** ets. in Maple). We will illustrate this with two examples



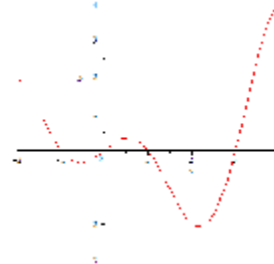
£

**Example 2.2**

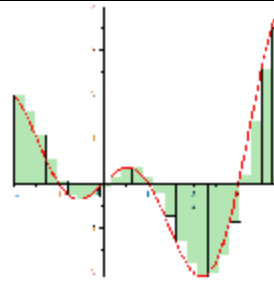
$$\int_{-2}^4 \left( x \cos \frac{\pi x}{2} \right) dx$$



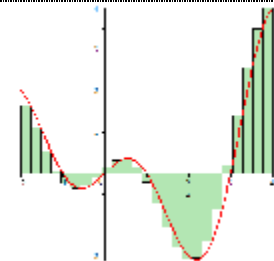
```
f := x -> x*cos(Pi*x/2);
with(plots):
implicitplot({y = f(x), x = -2, x = 4}, x = -2..4, y = -3..4);
```



```
with(student):
a := -2: b := 4: n := 20:
leftbox(f(x), x = a..b, n):
```



```
rightbox(f(x), x = a..b, n):
```



```
rightsum(f(x), x = a..b, n);
evalf(%);
```

$$\frac{6}{25} \sum_{i=1}^{25} \left( -2 + \frac{6}{25} i \right) \cos \left( \frac{1}{2} \pi \left( -2 + \frac{6}{25} i \right) \right)$$

1.060238075

£

The Maple's subpackage **Student[Calculus1]** has associated 17 Tutors. The **ApproximateIntTutor** command launches a tutor interface that computes, plots, and animates approximated definite integrals of a function on an interval . The command

*Student[Calculus1][ApproximateIntTutor(x\*cos(Pi\*x/2))]*

opens the dialog window as shown in figure 1. This Tutor application can be activated by command sequence **Tutors** → **Calculus** → **Single Variable** → **Riemann sums...** too.

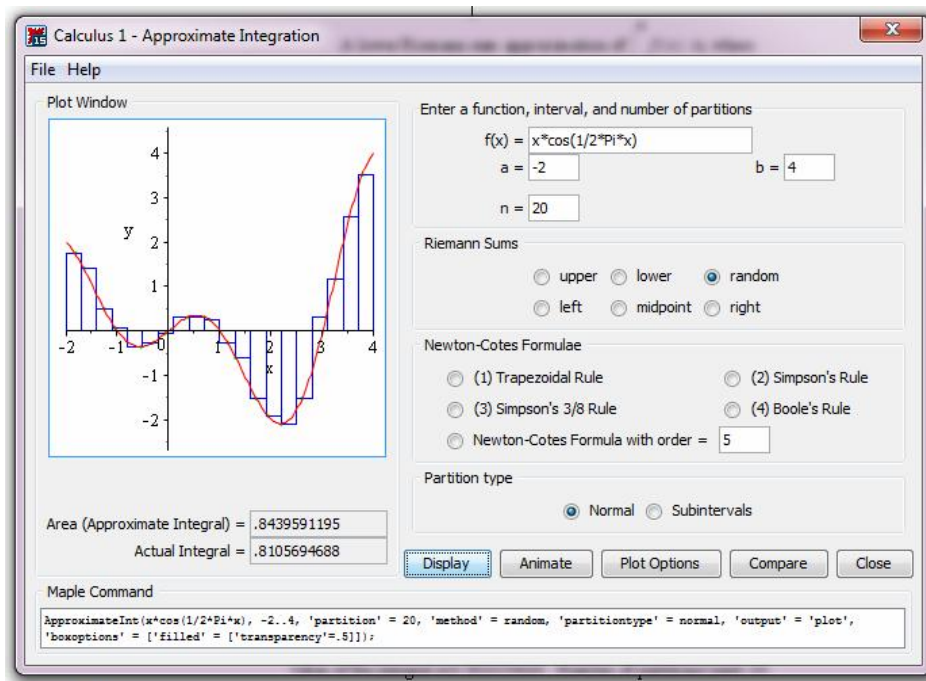




Figure 1

The Riemann sums tutor is a great tool to carry out computations with different Riemann sums. It also lets us visualize (with the help of animation) basic ideas behind the definition of definite integral.

### **Int() command and its “black box” effect**

One of the most important commands for computing (definite and indefinite) integrals in both CAS Maple and MATLAB /MuPAD is the command **int()**. The syntax for integrating an expression **expr** with respect to a variable **var** about indefinite integrals is **int(expr,var)** (note that Maple and MATLAB/MuPAD do not include the arbitrary constant of integration), and **int(expr,var=a..b)**, where **a** and **b** can be variables, numbers, or expressions. When students perform computations with command **int()**, the CAS tries to evaluate all expressions before returning the answer. In case the result is not as expected, it can be further processed by using commands like **eval()** and **simplify()**.



Example 3.1	Example 3.2
 $\int_1^6 \left( 10 \sin \frac{px}{2} - x^2 + 11x - 24 \right) dx$	 $\int_{-2}^4 \left( x \cos \frac{px}{2} \right) dx$
<pre>hold(int)(10*sin(PI*x)-x^2+11*x-24,         x=-2..4); = int(10*sin(PI*x)-x^2+11*x-24,       x=-2..4);</pre> $\int_{-2}^4 \left( 11x - 10 \sin \left( \frac{px}{2} \right) - x^2 - 24 \right) dx = \frac{-40}{p} - 102$	$\left[ \text{Int} \left( x \cdot \cos \left( \frac{\text{Pi} \cdot x}{2} \right), x = -2..4 \right) \right.$ $= \text{int} \left( x \cdot \sin \left( \frac{\text{Pi} \cdot x}{2} \right), x = -2..4 \right);$ $\int_{-2}^4 x \cos \left( \frac{1}{2} px \right) dx = \frac{8}{\pi^2}$

£

This way the students see only the final result and many of them become disinterested in learning how to integrate because the very process of calculating remains hidden. They often ask “Why should we be able to solve integrals when computers can not only do that instead of us but are also a lot faster?”. The use of CAS as “black box” (the term “black box” relative to computer algebra systems was first introduced by Buchberger 1989) to produce answers to mathematical questions with no attention to reasoning has received widespread criticism in mathematics education. Without knowledge of the underlying mathematics, many agree that the consequences of such use are disastrous for education and beyond [1], [2], [3] etc. The “black box” approach appears to add little to the educational experience as a whole [4].



### Use the First Fundamental Theorem of Calculus

A different way to compute definite integrals is to use its connections with the indefinite integrals via the Newton – Leibniz formula (also called “First Fundamental Theorem of Calculus”).

$$\int_a^b f(x) dx = F(b) - F(a),$$

where function  $F(x)$  is an antiderivative of  $f(x)$  (i.e.  $F'(x) = f(x)$ ).

To apply the first Fundamental Theorem of Calculus we use Maple or MATLAB to find an antiderivative function  $F(x)$  for the integrand  $f(x)$  and then we calculate the difference  $F(b) - F(a)$ .

<b>Example 4.1</b>  $\int_1^6 \left( 10 \sin \frac{px}{2} - x^2 + 11x - 24 \right) dx$	<b>Example 4.2</b>  $\int_{-2}^4 \left( x \cos \frac{px}{2} \right) dx$
The first thing to do is to define the integrand – function <b>f(x)</b>	
$f := x \rightarrow 10 * \sin(\text{PI} * x / 2) - x^2 + 11 * x - 24 ;$ $x \rightarrow 10 \sin \left( \frac{px}{2} \right) - x^2 + 11x - 24$	$> f := x \rightarrow x \cdot \cos \left( \frac{\text{Pi} \cdot x}{2} \right) ;$ $x \rightarrow x \cos \left( \frac{1}{2} px \right)$
Next, we try to evaluate one antiderivative function <b>F(x)</b>	
$F := \text{int}(f(x), x) ;$ $x \rightarrow \frac{11x^2}{2} - \frac{20 \cos \left( \frac{px}{2} \right)}{p} - 24x - \frac{x^3}{3}$	$> F := x \rightarrow \text{int}(f(x), x) ;$ $x \rightarrow \frac{2x \sin \left( \frac{px}{2} \right)}{p} - \frac{2 \cos \left( \frac{px}{2} \right)}{p^2}$
Finally we apply the Newton-Leibniz formula, i.e. we calculate <b>F(b)-F(a)</b>	
$(F _{x=4}) - (F _{x=-2}) ;$ $\frac{-40}{p} - 102$	$F(x) _{x=4} - F(x) _{x=-2} ;$ $\frac{8}{p^2}$

£

The process of integration is once again not apparent to the student but in this case the emphasis is placed on learning how to apply the first Fundamental Theorem of Calculus.

### Geometric Applications of Definite Integral

The definite integral is used for finding areas between curves, lengths of curves, volumes and surface area of three dimensional solids, and centroids of plane regions. By the lab-class assignments we can graphically illustrate and effectively solve a variety of problems from geometry.

**Example 5.** Considered the region R bounded by the graphs of the functions  $f(x) = x^3 - \arcsin(x)$  and  $g(x) = 3\sin(3x)$ .



- Find the area of the region R.
- Find the perimeter of R.
- Find the volume of the solid obtained by revolving the region R about the x axis.

We will use the following commands: *int()*, *numeric::fsolve()*, *assign()*, *plot::easy()*, *plot::Text2d()*. First we define the functions *f* and *g* in MATLAB/MuPAD.

Commands	Results
<code>reset():</code>	$x \rightarrow x^3 + \arcsin(x)$
<code>f := x → x^3 + arcsin(x);</code>	$x \rightarrow 3\sin(3x)$
<code>g := x → 3*sin(3*x);</code>	

Next we plot the graphs of *f* and *g*, using red for *f* and blue for *g*, for *x* in the interval from -1 to 1. We chose these numbers because the domain of *arcsin(x)* is the interval  $[-1;1]$ .

Commands	Results
<pre>A:=plot::easy(f,x=-1..1),   Colors=[RGB::Red]: Tf:=plot::Text2d("f = x^3+arcsin(x)",   [1,1.9]): B:=plot::easy(g,x=-1..1),   Colors=[RGB::Blue]: Tg:=plot::Text2d("g = 3sin(3x)",   [0.1,0.7]): plot(A,B,Tf,Tg,   ViewingBoxYRange=-3..3);</pre>	

Obviously, to answer any of the questions (a) – (c), the students have to find the points of intersection of the two graphs. The equation  $f(x) = g(x)$  cannot be solved algebraically and therefore when students try to solve it by using command *solve*, MuPAD returns no answer.

Commands	Results
<code>eq := f(x) = g(x);</code>	$\arcsin(x) + x^3 = 3\sin(3x)$
<code>solve(eq,x);</code>	$\text{solve}(\arcsin(x) + x^3 = 3\sin(3x), x)$

The students can obtain solutions of the equation by using the command **fsolve** included in the library **numeric**, that gives numerical answers, and allows us to specify the range in which we want a solution. Let us label the points of intersection as “a” and “b”, respectively, with “a” between -1 and 0 and “b” between 0 and 1.

Commands	Results
<code>numeric::fsolve(f(a)=g(a),a=-1..-0.8);</code> <code>assign(%):</code>	<code>[a = -0.8537067723]</code>
<code>numeric::fsolve(f(b)=g(b),b=0.8..1);</code> <code>assign(%):</code>	<code>[b = 0.8537067723]</code>

Not surprisingly  $a = -b$ . Observe that  $f(x)$  and  $g(x)$  are both odd. Hence, region R is symmetric with respect to y axis. With the points of intersection found, the students are prepared to set up the definite integral that represents the area of the region. We label the area “S”. The area S is equal to the sum of two areas  $S_1$  and  $S_2$ , where  $S_1$  is the area of the left part of the region between two considered graphics when  $x \in [a, 0]$  and  $S_2$  – the area of the right part of the region, when  $x \in [0, b]$ .

Commands	Results
<code>S1:=hold(int)(f(x)-g(x),x=a..0);</code> <code>eval(S1);</code>	$\int_{-0.8537067723}^0 (\arcsin(x) - 3\sin(x) + x^3) dx$ 1.309261482
<code>[S = 2*S1;</code>	2.618522964

Denote that the command **hold(int)()** display integral in standard printed form.

If the units on each of the axes are in centimeters, then the area of the region is 2.618522963 cm<sup>2</sup>.

To find the perimeter P of the region students have to add the arc length of the two pieces. They can use the familiar formula for the arc length to set up the appropriate integrals and then to evaluate by the command **float** the sum of two arc lengths.

Commands	Results
<pre> F:=sqrt(1+f'(x)^2): G:=sqrt(1+g'(x)^2): P1:=hold(int)(F,x=a..b); P2:=hold(int)(G,x=a..b);  eval(P1+P2); </pre>	$2 \int_0^{0.8537067723} \sqrt{\left(3x^2 + \frac{1}{\sqrt{1-x^2}}\right)^2 + 1} dx$ $2 \int_0^{0.8537067723} \sqrt{81 \cos(3x)^2 + 1} dx$ 12.75470765

The perimeter is approximately equal to 12.7547 cm.

To find the volume of revolution, we slice the solid perpendicularly to the  $x$  axis. The approximate volume of each “washer” is  $pg^2(x) - pf^2(x)$  for  $x \in [a, 0]$  and  $pg^2(x) - pf^2(x)$  for  $x \in [0, b]$ . The corresponding volume of the solid is then given by the integral.


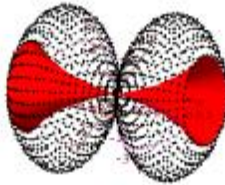
Commands	Results
<pre> F:=PI*(g(x)^2-f(x)^2): V1:=hold(int)(F,x=0..b);  V:=2*eval(V1); </pre>	$-p((\arcsin(x) + x^3)^2 - 9\sin(3x)^2)$ $\int_0^{0.8537067723} p((\arcsin(x) + x^3)^2 - 9\sin(3x)^2) dx$ 25.30210565

The volume of the solid obtained by revolving the region R about the  $x$  axis is approximately 25,30210565 cm<sup>3</sup>. £

In Maple to study geometric applications of definite integral, students can employ the Student [Calculus1] package and use commands as:

- **ArcLength()**,
  - **SurfaceOfRevolution()** and
  - **VolumeOfRevolution()**
- or the correspondence Tutors:
- **ArcLengthTutor**,
  - **SurfaceOfRevolutionTutor** and
  - **VolumeOfRevolutionTutor**.

We will illustrate some of these possibilities with the help of the next example.

<b>Example 6.</b> Find the volume of the solid obtained by revolving the region R bounded by the graphs of the functions $f(x) = x^3 - \arcsin(x)$ and $g(x) = 3\sin(3x)$ about the x axis.	
<pre>[&gt; with(Student[Calculus1]): &gt; f := x -&gt; x^3 + arcsin(x): &gt; g := x -&gt; 3*sin(3*x): &gt; a := fsolve(f(x) = g(x), x = 1);</pre>	 $a := 0.8537067723$
<pre>[&gt; VolumeOfRevolution(f(x), g(x), x = -a..a,   output = plot, axis = horizontal);</pre>	 <small>The solid of revolution created on <math>-0.8537067723 \leq x \leq 0.8537067723</math> by rotations of <math>f(x) = x^3 + \arcsin(x)</math> and <math>g(x) = 3\sin(3x)</math> about the axis <math>y=0</math>.</small>
<pre>[&gt; VolumeOfRevolution(f(x), g(x), x = -a..a,   output = integral, axis = horizontal);</pre>	$\int_{-0.8537067723}^{0.8537067723} \pi \left  (x^3 + \arcsin(x))^2 - 9\sin(3x)^2 \right  dx$
<pre>[&gt; VolumeOfRevolution(f(x), g(x), x = -a..a,   output = value, axis = horizontal);</pre>	$25.30210565$

£

The command

`Student[Calculus1][VolumeOfRevolutionTutor( $x^3 + \arcsin(x)$ ,  $3\sin(3x)$ ,  $x = -a..a$ )]`

opens the dialog window as shown in figure 2.

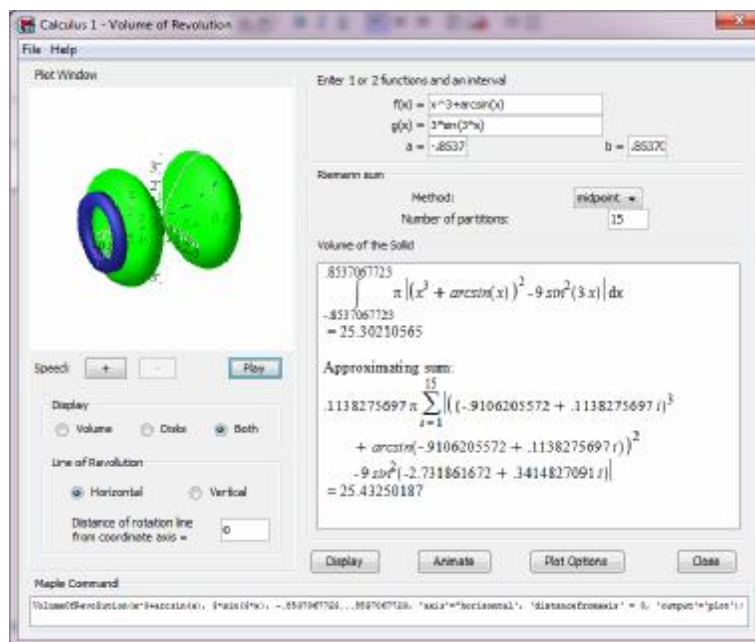


Figure 2

## Conclusion

Each lab-class assignment in higher mathematics has two main goals:

- To enrich the students' mathematical experience, and
- To show the students the variety of instruments and applications of the CAS they use.

The first goal is achieved through shifting the emphasis away from the drawing board to generating and applying creative ideas to find solution not only to complex problems that incorporate a lot of calculation, but also to unconventional problems that require profound understanding of mathematical theory as well as creative thinking.

The other goal is often considered not as important as the first one. Many lecturers tend to use a limited amount of commands to solve all kinds of problems, thus making the process monotonous and uninteresting, and lacking of any creativity. We believe that in the today's world, when people are surrounded by computers and digital gadgets, the ability to work with CAS will prove much more useful for young researchers than the skill to calculate by hand some not too complex integrals.

In our lab – classes, we constantly note that when the students do not have to face the hard task of calculating and transforming mathematical expressions, they manage to see the bigger picture and are able to understand the need of studying higher mathematics at university. Even students who encounter problems with mathematics, given the opportunity to use CAS, change their attitude to mathematics as a whole and succeed in solving problems which would otherwise invoke uncertainty and reluctance in them. Furthermore, the graphs are precise and well-drawn, which not only brings aesthetic delight, but also facilitates the understanding of the problem and makes the calculation easier and pleasant. This is particularly useful when solving different geometric applications of definite integral. To find interesting problems that incorporate spectacular graphs and unconventional solutions is an extremely challenging task for every successful professor.

### References

- [1] M. J. Bossé, N. R. Nandakumar (2004), Computer algebra systems, pedagogy and epistemology. *Mathematics and Computer Education*, 38(3), 298-306.
- [2] B. Buchberger (2002), Invention by lazy thinking, In: "A Min Tjoa @ Work"- In honor of the 50th birthday of Univ. Prof. Dr. A Min Tjoa, R. Wagner (ed.), Austrian Computer Society, 11-31.
- [3] M. A. H. MacCallum (2002), Computer algebra in general relativity, *Int. J. Mod. Phys. A* 17, Special issue: Proceedings of the conference on general relativity, cosmology and relativistic astrophysics (Journées Relativistes), ed. C. Barrabès and P.A. Hogan, 2707-2710.
- [4] P. Drijvers (2000), Students encountering obstacles using a CAS. *International Journal of Computers for Mathematical Learning* 5(3), 189 - 209.

### Authors:

**Zlatka Mateva**, PhD, Department of Mathematics and Foreign Languages, Section of Mathematics, Technical University of Varna, Studentska Str. 1, 9010, tel. 052/383 228, e-mail: [ziz@abv.bg](mailto:ziz@abv.bg)

**Nedka Pulova**, PhD, Department of Mathematics and Foreign Languages, Section of Mathematics, Technical University of Varna, Studentska Str. 1, 9010, tel. 052/383 250, e-mail: [vpulov@yahoo.com](mailto:vpulov@yahoo.com)

**Violeta Bojikova**, PhD, Department of Computer Science and Engineering, Technical University of Varna, Studentska Str. 1, 9010, tel. 052/383 616, e-mail: [vbojikova2000@yahoo.com](mailto:vbojikova2000@yahoo.com)



## AGAINST THE BOREDOM TABLET. GET ONE!

Desislava Milenkova, Yordanka Parvanova

**Abstract:** *We present one more different way to education. The new one. The education of 21<sup>st</sup> century. That, what the nowadays kids need. Education, which is connected to their necessities, which teaches them how to orientate in the informational stream of knowledge.*

**Keywords:** *educational games, interactivity, e-Learning*

World of technology. This exciting world, through which we learn so much every day, share, communicate and entertain ourselves. This is a world where in order to check the time we look at our phones.

Sometimes we have a hard time accepting the digital 21st century. But this is the world of the children and the new generations..

The education should always work in the best interest of the children and not against them.

Children today have unlimited quantity of resources and information. But do they know how to use it? Do they develop the ability to distinct the important from the unimportant information? The teacher's goal should be helping the students with their interaction with the world, to guide them to variety of ways to extract, synthesize, analyze and use the information at hand, because "Flipping a boring lecture from the classroom to the screen of a mobile device [is] the same dehumanizing chatter, just wrapped up in fancy clothing." [1]

The classroom can turn into a spring of fun learning. In fact the classroom can go out of the rules that are set before. We can turn the world into a place where children can learn, solve problems, to explore and discover the information they need.

The real world is the best classroom - the place where learning continues throughout your whole life. And learning through "doing", through increasing the activity of the students is far more useful than the usual sitting in the classroom where there is no dialogue between the teacher and the students.

What is the right application of the digital technologies in this different environment?

One of the pros of using educational software is to help the teacher translate their knowledge better and that helps them immensely. Products like Envision and Jumpido change the established statu quo and challenges the students to be part of the educational process.

It's a fact, that students are in constant interaction with the technologies – desktop and mobile systems - this is their reality. This challenges the teachers to search for their application in their own

educational process not only as a mean of entertainment, but as a way of expanding the skills that are needed in the 21st century.

This has led the authors of this article to the usage of the mobile devices in their work. The main challenge is to expand the educational environment outside the classroom. While that's exciting and innovating, it's also needed to be administrated, customized for the needs of the educational process and a way of them to be transferred in the children's personal devices. "If you're not prepared to be wrong, you'll never come up with anything original"[2], so that we are not afraid to experiment in education. "Educational technology is the study and ethical practice of facilitating learning and improving performance by creating, using, and managing appropriate technological processes and resources." [3]

In the modern world there is wide spread usage of the word "connect" (literally and metaphorically) with the coming of the mobile devices. These smart devices have one really useful (sometimes) function - when we write, it recognizes and complete words - it's a time-saver, but it's often occurrence that those words are inappropriate or

Strange. One of those completions suggests that "tablet" is "tablet" (which is homophone for a pill in Bulgarian). Is this acceptable use of homophone in this boring and old way of education? There are great examples that show that the education is one step behind and requires a new image in the modern world. It's just like going to the cardiologist's office then to a classroom and find the the difference 100 years made.

It's not hard to imagine the modern teacher in a classroom 100 years ago. The chalkboard might be different, maybe not.

Our challenge is to change the image of the educational system by seizing the future, because the future - that's our children.

#### References

- [1] Musallam. R, [http://www.ted.com/speakers/ramsey\\_musallam](http://www.ted.com/speakers/ramsey_musallam)
- [2] Robinson, Sir Ken, [http://www.ted.com/speakers/sir\\_ken\\_robinson](http://www.ted.com/speakers/sir_ken_robinson)
- [3] Educational Technology at the University of Hawaii'i, <http://etec.hawaii.edu/aboutbackground.html>

#### Authors:

**Desislava Hristova Milenkova**, Secondary school "Angel Kantchev", Sofia, Bulgaria, 0896 805 596, [kidsbg@gmail.com](mailto:kidsbg@gmail.com)

**Yordanka Parvanova Parvanova**, Secondary school "Luben Karavelov", Dimitrovgrad, Bulgaria, 0886 803 499, [dany.parvanova@gmail.com](mailto:dany.parvanova@gmail.com)

## M-LEARNING IN EDUCATION

Mitko Mitev, Ivaylo Penev

**Abstract:** *Mobile learning (m-learning) is a contemporary trend in education, concerning all its degrees, which affects not only the used mobile communication tools, but has also direct reflection to methodology and learning process organization. In the paper the following essential problems are concerned:*

- *The provided hardware and programming possibilities of mobile devices for communication,*
- *Instrumentation tools for programming interface design, concerning specifics and requirements of learning process,*
- *Approach for development of courses, based on mobile learning,*
- *The advantages and disadvantages of m-learning are discussed,*
- *Definite recommendations, concerning m-learning in different areas of education are made.*

**Keywords:** *m-learning, text metrics, ontology, finite graphs, causal links, learning units, learning groups, clustering, input/output links, learning control, testing, individual learning*

### Introduction

Education is conservative in nature, but it is also a subject of intensive development, which is based on the following assumptions:

- Education is a reflection of the overall development of society.
- The main goal of education is to support this development in current and future stage.
- Education lays in the basis of human nature of individual, his desire for development in a particular social environment, natural curiosity, willingness to learn the content and meaning of the processes, phenomena, relationships, etc., taking part in society and in nature.

Natural development of society, assisted by education, has an impact on the educational process by imposing its demands on the used methods, tools, forms, and on the training period.

The conservative components of education are:

- Basic facts and knowledge, which are the basis of existing world and therefore do not change. Learning their nature only enriches our knowledge of them. Based on them are carried out various experiments and tests, and accordingly scientific theories are developed, which aim not only scientific explanation of the current facts and events, but also their purposeful "use " to achieve the desired objectives.

- In the learning process there are always two components: a source and a receiver of knowledge and skills, i.e. an exchange of information with different intensity exists, which is often mixed, not always effective and in certain cases even negative, according to the goals defined.

### 1. Informational Resources

In historical aspect, the main source is the teacher. With his experience and knowledge, on the basis of his methodological skills, he has organized, carried out teaching and has registered the achieved results. Currently, teacher is one of the main, but not the only source of information and educational impact on the benefit of learners.

A relatively generalized classification of information sources is as follows:

- Textbooks - classical form based on purposefully selected, summarized and methodologically processed information, provided to students in a different way in a certain medium (paper, magnetic, optical) or by means of communication, verbally, visually, etc.

- Regulated information sources for basic and further learning, which in form and content are similar or are completely covered with the front group of sources. In addition to this group belong the scientific literature: books, monographs, articles and reports from scientific and experimental research, regardless of the medium used.

- Communication and information interaction between teachers and students and between the participants in each group themselves. The process of transmitting and receiving information can be regulated (classical forms of education: lectures, exercises, essays, course works, course and diploma projects, etc.), free discussions, alone or under the guidance and leadership of teacher, exchange of knowledge and experience in the collective performance of a given task as a component of the learning process and more.

- Research, development and documentation of scientific and applied problems as a part of the learning process by assigning additional or self-developments on the basis of individual interest, etc.

- Social networks (Facebook, YouTube, Twitter, LinkedIn, Wikipedia, Google, created and maintained common or thematic blogs, etc.). The communication and information possibilities of the listed social networks are practically limitless: transmission and sharing of text, video and audio information, exchange of thoughts, opinions and reviews, freely organizing events of a different nature and more.

- Specialized Networks (NREN-National Research and Education Network, UNICOM - Bulgarian branch of NREN), ResearchGate for recording and exchange of scientific information and more.

- Mass Media (TV, radio, cinemas, theaters, concerts, etc.).

These are just some of the main sources of information, on the basis of which are prepared curricula, programs. They are methodologically summarized and processed for the purposes of learning, etc.

Each of these sources should be reasonably examined on the basis of preliminary defined objectives and to examine their opportunities in the

aspects of communication, awareness, methodological readiness, technical and program implementation for the purposes of education and particularly for mobile learning. In this respect the published sources can be explored depending on their nature, significance, and in accordance with the stage of development of training courses:

- ü Text metric studies [7, 8] provided by text information sources. The results of these studies provide data for the frequency of occurrence of certain terms used and the combinations between them. The data significance is indirect, but has a specific meaning in graph analytical approach to development courses.

- ü Ontological studies [6, 9] mainly reflecting to some extent the semantic content of the terms used, their connectivity, classification and structure. Developments are related to specific areas of application. In [17] are proposed tools for the creation of the so called ontological classes, which particular application could be used in the research process and subsequently to create software for learning purposes.

- ü Investigation of causal [3, 4, 12] links between objects, processes, facts, concepts, etc. The results of these studies can be used in structuring the learning material in chronological arrangement of the identified learning units, their grouping by volume and content and other design decisions, related to the development of courses and programs.

## 2. Hardware and software tools for mobile learning [2, 18].

M-learning as a concept and form of organization and training is considered [19] as a specific form of e-learning and distance learning component. Essential meaning for m-learning have the technical means and the software to effectively conduct the learning process. The range of these training tools is practically unlimited. For the purposes of mobile learning this range is reduced to devices such as smartphones, tablets, gaming consoles, and others. Regardless of the brand and a particular manufacturer, their functionalities are well known. In general, they could be the following:

- ü Establish and maintain a certain communication between the source and the receiver for transfer of information [1]. The main problems are technical realization and maintenance of certain channels, signal strength, connection reliability, the possibility of a sudden collapse of the established contact, punctuality (geo location) interference (call, send messages, SMS, MMS, etc.), internal (user and/or program interruption, passing in the background, low battery, etc.). The impact of the above problematic situations is associated with disruption of the learning process. In most cases, the communication can be restored or continued without substantial loss of information. When developing software for

mobile learning these possible situations must be reported and their process must be included in the project solution.

- ü Technical and design parameters of the mobile device. The main are: the size and resolution of the monitor volume of available memory for mobile application (the so called isolated memory, battery capacity, maintained color palette, etc.),

- ü Completing the device including camera/video camera, accelerometer, coordination, status, etc.

- ü Software: operating system (Android, iOS, Windows Phone, etc.), availability or access to libraries, software tools, including additional programs, etc. This includes programs for implementation, maintenance and reproduction of sound and picture, geo location , positioning and orientation, etc.

In application aspect, the capabilities of mobile devices for communication and exchange of information for the organization and conduct of the learning process are partially addressed in [3, 15].

### 3. Organization and forms of e-learning

Different organization structures for providing and conducting mobile learning can be classified as follows:

- ü Centralized, classical structure of traditional learning. The advantages are well known. Some of them are: collective organization, control and monitoring of the learning process updated for all visual contact between teachers and students, documentation, etc. Disadvantages: not reported individual capabilities and current students, the dominance of the conservative element is valid for all, etc.

- ü Decentralized alternative to the classic structure. The advantages are mainly related to the autonomy of learners units dynamics applied forms of learning, diversity in teaching material conditions of competition and others. The main disadvantage is that they can produce significant deviations from the general requirements for learning outcomes as a result of subjective views or reporting of cyclical conditions.

- ü Fixed governance structure defined by relatively resistant controls, localized in specific sites for centralized or decentralized learning process.

- ü Partial mobile structure. To this group belong organized and maintained forms of learning after completing a certain level of education that an individual or collective nature, are regulated in content and properly documented. This structure allows eligibility to the direction and subject area, depending on the specific requirements of individual realization.

ü Mobile structure. It belongs to the so-called "Learning for life" in which the learning process is not regulated in form and content, and form the basis of various factors such as changes in the environment and production environment, individual memorial, need for natural development and others. Forms of education are diverse and range from regulated to arbitrary independent.

#### 4. Preparation and conduction of mobile learning courses

In the development of training courses the following classic stages are performed:

ü Clarification of goals and tasks of the developed course.

ü Definition of the subject area

ü Determination of the course in accordance with the objectives and tasks

ü Establishment of the so called interdisciplinary links with businesses in the subject area, not included in the course (immediate, direct connection) and outside the area (away, indirect connectivity)

ü Identification of information sources that support the content of the resulting course, they can be classified as basic, additional and auxiliary, grouping them according to content, form and manner of presentation, communication and access.

ü Development of curriculum and / or program.

Development process is dynamic and iterative for reasons related to changing conditions and requirements, the natural evolution in the development process, the emergence and establishment of new facts, phenomena, processes and other objective and subjective factors that are directly related to the nature and content itself process.

In the development of mobile learning courses the specific features and capabilities of mobile devices for organizing and conducting courses, should be taken into account, namely as follows:

ü Training is not associated with a place for learning process.

ü Learning process is not associated with regulated time.

ü Training is individual and specific to each student.

To satisfy these requirements the following technological sequence for designing and conducting mobile course is recommended:

4.1. Identification of the smallest units of learning that encapsulate themselves in a relatively independent information about an event, a process, phenomenon, etc., or separate parts.

4.2. Establishing causal links between them. The link is directed from the base unit to the learning unit, which is based on the previous one. The previous one, in turn, could be base for the next one and so on. This generalizes the rule that the base unit must precede its sequence.

4.3. Methodologically properly structured training course requires all educational units to be arranged along the axis of time, so that the

basic units precede their sequences. In general, this requirement cannot be satisfied entirely by the complexity of the structure or cross- couple units in which each is both base and effect of the other.

4.4. Rational decision is those one, in which the majority of the causal links are properly available and the "distance" between the basic concept and its sequence is minimal.

4.5. Similarly causal relationships between teaching units in previous or subsequent courses are identified and evaluated. This indicates the influence on the arrangement of training units for the time axis.

4.6. Clustering of learning units in study groups regulating the volume and content of training material, which in a moment of time is granted and requires learning. The criterion is also correct positioning of the causal relationships between groups to minimize their number and consequently rearranging educational units within each group. This approach allows individual grouping of educational units, according to the specific capabilities of the student and its current status.

4.7. As a result of control (exam or test) carried out may be established training units, which are not learnt sufficiently and should be learnt again. Therefore, these units can be rearranged in accordance with the causal links between them and their connectivity with other academic units.

For the realization of this technological sequence are necessary formal description, development of a set of algorithms for finding rational solutions and their program implementation as instrumental tools for developing mobile learning courses and dynamically adaptable management of the learning process.

The formalization of the problem in [2,12] is reduced to a finite directed graph (fig. 2) with weights of vertices (learning units) and the static and dynamic weight of arcs (causal links). An algorithm is developed for content structuring [4], clustering of learning units [5] and thus determining the composition of the starter kit of concepts [11], decomposition of the learning units in groups [14], algorithms related to the management of mobile learning [10,15]. The main part of the algorithms are implemented [13,16].



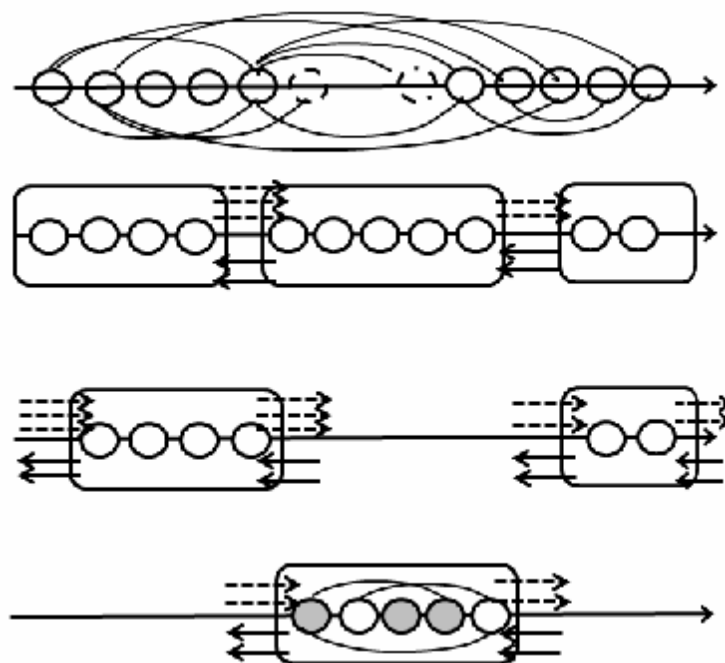


Fig. 1. Finite directed graph of learning material

#### Conclusion

The problem for the application of mobile learning in education is updated as a result of the development of society, new requirements regarding the form of learning, the need for lifelong learning and as an objective prerequisite the emergence of new mobile communication devices, development of sources of information, social and academia, etc.

For this purpose it is necessary to study the characteristics of information sources thoroughly and professionally, as well as the methods and means of communication devices for presentation of text, audio and video information, new ways of social interaction and this issue in the context of their application in mobile learning courses.

#### References

1. Stefanov. D. *Mobile learning from home*. Vol. "Modelling and management of information processes", KTP, Sofia, 2009, ISBN: 978-954-9332-55-1, 65-69 (in Bulgarian language)
2. Mitev M., E. Racheva, G. Marinova, *Graph analytical approach for development of distant learning courses*. Fourth international conference "New

information technologies in education for all: innovative methods and models", Kiev, Ukraine, 24 - 26 November 2009 г., pp.127-136 (in Russian language)

3. Mitev M. I. Stoev, Application of mobile devices in learning., Third international scientific congress, Technical University - Varna, Vol.1,129-134, 2012 (in Bulgarian language)

4. Mitev M., E. Racheva, *Algorithm for initial definition of notions structure in distance learning*, I International scientific-application conference "System Analysis. Informatics. Management", Zaporozhie, Ukraine, 4-5 March 2010., Conference Proceedings, pp. 127-128 (in Russian language)

5. Mitev M., E. Racheva, *Algorithm for initial grouping of notions in distance learning*, National Academy of Sciences of Ukraine, Institute of Cybernetics V. M. Glushkov, International Scientific Journal «Management Systems and Machines», № 2, March-April 2010 г., pp. 13-16, UDK 681.3:658.56 (in Russian language)

6. Mitev M., E. Racheva, *Ontology and Онтология и дистанционное обучение, обслуживающее промышленное производство*, VI International Conference "Strategy of quality in industry and education" (4-11 June 2010, Varna, Bulgaria): Proceedings in 4 volumes. Vol. II (P. 2). Editors: Hohlov T.C., Hohlov B.A., Stupak U. A. – Dnepropetrovsk-Varna, 2010. (in Russian language)

7. Mitev M., E. Рачева, *Text-metric approach for preparation of distance learning courses*, V International Conference ITEA-2010, "New information technologies in education for all", 23-25 November 2010, Kiev, Ukraine (in Russian language)

8. Mitev M., E. Racheva, D. Dimitrov, *Application of text-metric research in development of specialized interpreters*, Ukraine National Academy of Sciences, Institute of Cybernetics. V. M. Glushkov, International Scientific Journal «Managing Systems and Machines», March-April 2011, UDK 681.3:658.56 pp. 78-81 (in Russian language)

9. Racheva E., M. Mitev, *Ontological approach to increase the quality of distant learning*, VIII International Conference „Strategy of quality in industry and education ", 8 – 15 July 2012, Varna, Bulgaria, pp. 304-306 (in Russian language)

10. Mitev M., E. Racheva, *Major Challenges to the m-learning Organization*, ITEA'2012, Kiev, Ukrain, 27-28 November 2012

11. Mitev M, E. Racheva, G. Marinova, *An Algorithm for Defining of Initial Package of Notions for Purpose of E-Learning*. Proceedings of the International Conference on Computer Systems and Technologies - CompSysTech'09, 17-18 June, Ruse, Bulgaria, pp. IV.9-1- IV.9-5

12. Mitev M., E. Racheva, *Graph-Analytical Model for Concepts in E-learning Content*, International Scientific Conference UNITECH'11, Gabrovo, Bulgaria, 18-19 November 2011, p.p. III-311 – III-315

13. Mitev, M., A. Antonov, I. Penev. *An algorithm and program for management and evaluation of individual distance learning*. Proceedings of the 6th International Scientific Conference "eLearning and Software for Education", Editura Universitara,, Bucharest, Romania, pp. 261-266.

14. Mitev, M., A. Antonov, I. Penev. *Algorithm for decomposition of learning content*. Proceedings of the 8th International Scientific Conference

“eLearning and Software for Education”, ISSN: 2066-026X, Editura Universitara, Bucharest, 2012, pp. 212-217.

15. Mitev, M., A. Antonov, I. Penev. *M-learning in engineering education*. Proceedings of the 9th International Scientific Conference “eLearning and Software for Education” vol. 2, ISSN: 2066 – 026X print 2066 – 8821 online, Editura Universitatii Nationale de Aparare “Carol I”, 2013, pp. 183-187.

16. Mitev M. D. Georgiev – Program Library for Development and Research of Distance Learning Courses. XVI International Scientific conference on information, communication and energy system and technologies ICEST 2010 Ohrid, Macedonia

17. Nikolov S., Antonov A., *Framework for building ontology-based dynamic applications*, In Proc. of CompSysTech '10, 2010, pp.83-88

18. E-Learning Resources, m-Learning (mlearning) - <http://www.grayharriman.com/mlearning.htm>

19. Crompton H.- *A historical overview of mobile learning: Toward learner-centered education*. Handbook of mobile learning. Florence, KY: Routledge. 2013, 3-14 - [http://www.academia.edu/5601076/A\\_historical\\_overview\\_of\\_mobile\\_learning\\_Toward\\_learner-centered\\_education](http://www.academia.edu/5601076/A_historical_overview_of_mobile_learning_Toward_learner-centered_education)

**Authors:**

**Mitko Mitev**, PhD Eng., Assoc. Prof. TU-Varna,  
052 383 604, e-mail: [mitevmm@abv.bg](mailto:mitevmm@abv.bg)

**Ivaylo Penev**, PhD Eng., Assistant TU-Varna  
052 383 409, e-mail: [ivailopenev@yahoo.com](mailto:ivailopenev@yahoo.com)

## INTERNET AND AGGRESSIVE BEHAVIOR

Joanna Nedyalkova

**Abstract:** *One of the typical behavioral reactions in humans is aggression. The invention of the Internet is not only not conducive to humanity in its strict restriction on the contrary, creates and promotes a number of new opportunities for the development of different for of aggression.*

**Keywords:** *Internet, aggressive behavior*

Aggressive behavior is one of the main topics of modern psychology and it is a part of our lives. Each of us in certain situations behave aggressively in its own way. Aggression develops in the family, improve in school and became a companion in our lives. It is not possible to go into the nature of such behavior without understanding the reasons that cause it. According to J. Zografova, aggressiveness is instrumental function for active adaptation of the individual to the human environment and as such is of interest to social psychology. "This is a personal behavioral strategy to achieve different goals of the person it is a major part of its relationships with other people." J. Zografova problems of aggression in humans, S. 1987, p 29-30 /

Today aggressive behavior often times is an example of a winning strategy and the willings to follow or repeat are too much, especially teenagers. Many are writing on the subject, and in fact does so little. Approaching the problem of the conditions for aggressive behavior, it should be noted that like any human behavior and aggressive is determined by the interaction of two factors: personality and situation. Situational conditions are presented in two groups:

- 1/such as acting over a long period before the aggressive behavior;
- 2/going immediately prior aggressive behavior.

The issue of aggressive behavior is examined in medicine, biology and others sciences, but one of the most convincing results are in the field of psychology of personality. According to the dictionary of terms psychology, aggression, hostility is a trait of the individual, stressing its tendency to cause trouble, to attack, to cause injury to other people and the world around. The problem of aggression and self-aggression emerged in focus whenever there is a disruption of social communication when violated human values and life. It is important to emphasize that according to scientists forms of aggression are different and it depends on the cultural environment in which we live. Personal characteristics of the people further complicate the understanding of aggressive behavior and predict when it will occur. In Bulgarian dictionary aggression is a behavior aimed at eliminating or overcoming any obstacle or threat striving assault, conquest, plunder. / Sofia Science and Art, 2005 / The main variables that affect aggression at the individual level are the very

personality and its attitudes. This behavior should be considered inextricably linked with the explanation of anxiety, anger and violence as emotional states. There are differences between the listed events, but in practice and everyday life is difficult to distinguish them, especially between aggression and violence, because a person's motivation to them there may be complex. Television programs, the impact of media and internet on the human psyche is a problem, often discussed especially nowadays, from both journalists and scientists working in the field of psychology and sociology. Internet defines many of the minds of teenagers and plays a major role in their identity. While to the elders, Internet is a relatively new invention for kids today it is part of their daily lives and the real world, because it was there when they were born. For them it is something as natural as the phone and the TV, but very attractive, because they participate in what is happening there. Most of the teenagers are on the Internet because he gave an important opportunity to communicate. It ranks first among all classes of children on the Internet. They communicate with their friends and peers both in real life and out through the Internet and do not distinguish between the two forms became one for the other. They learn many things from the Internet, create in it, upload photos, which they did, share ideas, and discuss. But like physical world and the Internet is fraught with danger and risks that children should be taught by small, but it seems increasingly still insufficient work on the problem. The scenes of violence, aggression and blatant violation of generally accepted values may provoke an already begun a process of social deviation. Through images on the internet games world that begins to be perceived as hostile and life with others filled with threats and violence. We can definitely say that computer games aggression occurs most commonly in the rudest form. The main issue that is discussed is whether the demonstrated violence causes aggressive behavior. After an experiment, Bandura and his colleagues concluded that the observed aggression causes such behavior and in the viewer. Especially children copy the behavior of adults. The results of the study show that the majority of students believe that sending aggressive emails is mostly just for fun and only in few cases the aim is to hurt the other. The study demonstrated that most children with aggressive behavior on the Internet do not deliberately willing to harm others. Even when messages are insulting or demeaning both sides accept them as a joke. Seems to have originated such a culture in the space of communication through hatred and aggression. The adoption of this form of behavior unfortunately can be explained by the lack of in-depth treatment by many adults and online violence goes online bullying and strong aggression. Most often verbal aggression is accompanied by insults, have an additional role as punctuation marks like exclamation marks and also writing in capital letter. This type of

aggression is called "flaming" / flash / and occur mostly in places of discussion and exchange written comments such as blogs, forums and etc. Interesting to note that the exchange of angry insults and messages is aggression not only on the individual himself against which focuses on, but also to others, passive witnesses of what happens. The wave of anger triggers and violates both the tone and theme. So comes the end of many discussions. The written form is not the only way to cause resentment and anger. With the development of the technological aspects of the virtual presence of many emoticons graphics allow for the conduct of wars between Internet users. Another typical form of Internet aggression is sexual violence. Behind the mask of anonymity, it becomes a free field for the expression of repressed desires. Virtual environment and the freedom that gives the majority of introverts in real life is a prerequisite for numerous violations of the individual and his psyche. A most favorable environment for this type of aggression are dating sites.

Internet aggression is good ground for expression and personal correspondence of consumers. The mere existence of "open" mailbox is a prerequisite of abuse the holder. The reasons for Internet aggression are extremely numerous and inhomogeneous. This puts researchers to inability to define clearly and precisely and rapidly changing hinder them to identify unequivocally.

Unfortunately the victim of aggressive behavior can be anyone. Internet and the physical world hold risks and dangers, that children should be trained to avoid from small. For the youngest, parents should to use the ability to block inappropriate content or install a special program for parental control. But - it is important to talk to children about the Internet to learn to share what bothering them. Should work to take effective measures to change existing attitudes through the children in non-violent communication and social skills, to promote positive social norms in cyberspace.

### **References**

1. Gercheva-Nestorova, G. Social Communication. VSU University Press, 2012
2. Zografova, I. Problems of aggressiveness in humans. S., 1987
3. Zogarafova, I. Investigation of aggressive behavior in terms of social change. S., 2001
4. Certain explanations for aggressive behavior, / [www. e-placebo.com](http://www.e-placebo.com) /, 2014
5. Stankov al. Internet aggression

### **Author:**

**Joanna Mihailova Nedyalkova**, PhD, Varna Free University

## STACKS AND QUEUES IN DEVELOPING CARD AND MEMO GAMES

Eleonora Pavlova

**Abstract:** *Today we experience an enormous expansion of computer technology. This rapid growth has left a big gap in the supply of developers around the world. As educators we have the opportunity to do something new, something powered by the high-tech for the high-tech – interactive teaching. Scratch is fantastic interactive introductory tool that makes it possible to teach computer programming from early ages in a fun, save environment. Scratch helps young people learn to think reason systematically, to understand how algorithms and data structures work together to produce the correct output.*

**Keywords:** *stack, queue, data structure, interactive teaching, algorithm, computer programming, memory game, Scratch*

Nowadays we experience an enormous expansion of computer technology. Just think about the millions of applications for the millions of smartphones and tablets which provide us with different services, education and entertainment all around the world. This computer revolution has left a big gap in the supply of developers around the world - we need more people in technology and we need them ASAP.[2]

Interactivity is what makes the IT so much more than just calculations. It provides entertainment, challenge and skill development in an easily accessible package. However, the focus is not on helping adults or university students to become interested in programming but rather on building a passion for the younger audience. The spotlight is on kids in the age group 7-16. Teachers cannot start with the cumbersome academia or lengthy code writing – a child would not be interested in that at all.

Educators knew a new way to catch these children's attention was needed and they have come up with a solution. Learn and play at the same time. Give any kid a game and you would instantly see how they are now ready for the challenge, truly eager to play along. By using this method, we can provide inspiration for the younger generation, to become hungry for more in the field of computer technologies. This is the way to help these kids become the striving younger and ambitious people, entering the IT sector with innovative ideas and vision. As educators we have the opportunity to do something new, something powered by the high-tech for the high-tech – interactive teaching.

Scratch is a programming language, designed with learning and education in mind. With Scratch, students can program interactive stories, games, and animations using their own imagination, logical thinking skills and creativity. This is fantastic interactive introductory tool that makes it possible to teach computer programming from early ages in

a fun, save environment. Scratch helps young people learn to think reason systematically and to understand computing concepts - essential skills for life in the 21st century. Scratch is a project of the Lifelong Kindergarten Group at the MIT Media Lab.[4]

Games have been used at all levels of learning from elementary to college. Children like to play and they are interested in learning how to develop computer games. On the other hand, game development is one of the fastest growing areas of software development. "A game-based assignment will typically contain several components such as graphical user interface (GUI), data manipulation, file management and algorithm design. A completed computer game reflects a significant understanding of all the algorithms and concepts needed. Other important aspect of game implementation is the selection of suitable data structures."[1]

Understanding how algorithms and data structures work together to produce the correct output is a skill that our young students ought to possess. It is also important to realize that not all data structures or algorithms have a specific implementation. One should always analyze exactly what are requirements of the game and then select the appropriate algorithms and data structures. The most important data structures for game programming are Binary search tree (BST), stack and queues. "Stacks and queues should be implemented in Scratch with list variables. A list can hold several values at the same time".[5]



Figure 1. List blocks in Scratch

At first the students solve the problems without prior knowledge of data structures, although their intuitive solutions use them. The first tasks given are selected in order to aid the teacher in explaining the purpose and features of stacks and queues. For example, the queue tasks may include putting in 10 random numbers from 1 to 10 which are later shown in the output in the same order as in the input. In order to explain how



stacks work, we use the same task, but the numbers in the output must be in the reverse order.

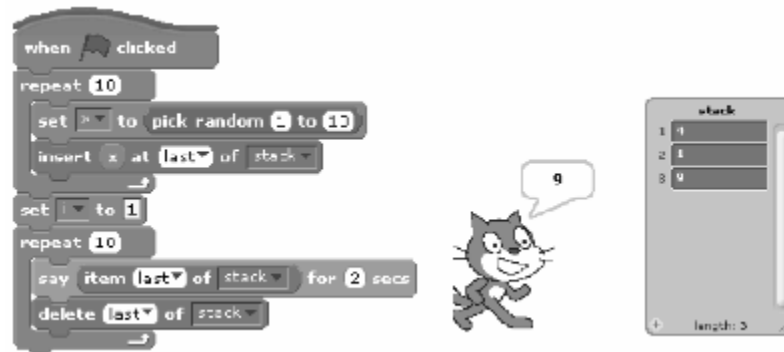


Figure 2. Stack – list implementation in Scratch

The code shown in Figure 2 will put in 10 random numbers from 1 to 10 in the list using the command: **insert x at last of stack**. The next number will appear as element 4 in the list. After that the output starts from the top of the stack with the command: **delete last of stack** – i.e. the last number that was put in will be deleted first.



Figure 3. Queue – list implementation in Scratch

10 numbers from 1 to 10 chosen at random are put automatically in the queue with the command: **insert x at last of queue**. The output starts with the first number put in the queue with the command: **delete 1 of queue** (Figure 3).

Data structures stack and queue are often used in card and memory games in order “to hold” the cards or to save the sequences of numbers, colors, sounds and ets. Simon Game is a memory game

traditionally known in the form of toys. The computer version of the game fulfills the same objective - to challenge an in a way measure the memory retention capacity by generating a growing sequence of events - colors and sounds, that the player has to repeat. In order to create the program in Scratch we need a list with colors that the player must repeat. We named that list **list\_color**. The data structure queue is used to realize the color sequence. Every new color in the task sequence as an element of the list\_color goes at the end of the queue (Figure 4). When the player clicks the color correctly, we have to delete the first element of the list\_color (Figure 5). Queue is known as a FIFO structure (First In, First Out). The element that gets into the line first will be checked out first.



Figure 4. Simon\_game adding element in queue "list\_color"

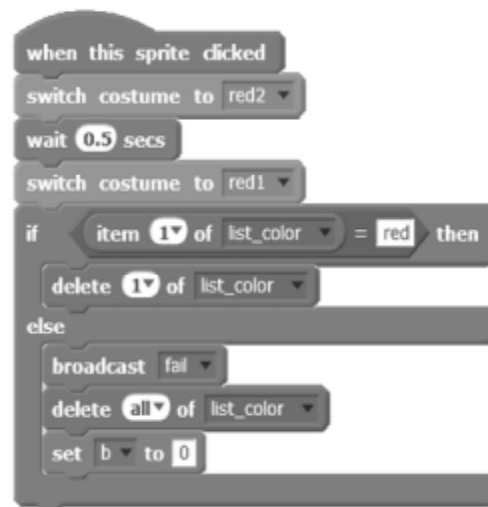


Figure 5. Simon\_game deleting an element from queue "list\_color"

The "War" is a card game for two players. The deck is divided among the players giving each a down stack. Players reveal the top card of their deck - this is a "battle". The player with the higher card takes both the cards played and moves them to the bottom of their stack. If the two cards played are of equal value, then there is a "war". Players draw two cards face-down and another card face-up. The owner of the higher face-up card wins the war and adds all six cards on the table to the bottom of their deck.

The deck of each player realizes with a queue – **deck1** and **deck2** (Figure 6). The lists **war1** and **war2** hold cards that were drawn during a war. The **war** lists are stacks – the last drawn card has to be faced up. The relation between these last cards in stacks **war1** and **war2** determines the winner in the war (Figure 7).

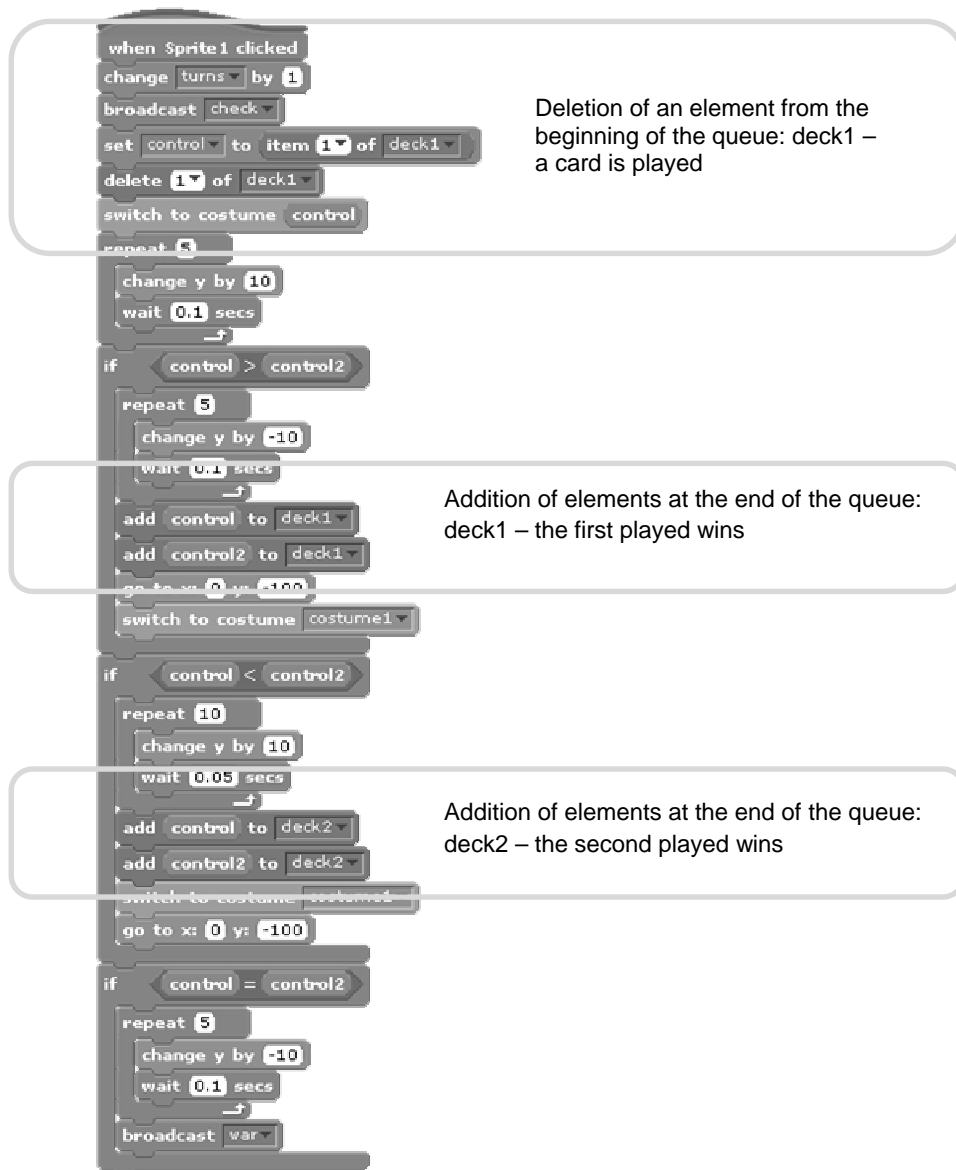


Figure 6. Card game "War" – "Battle"

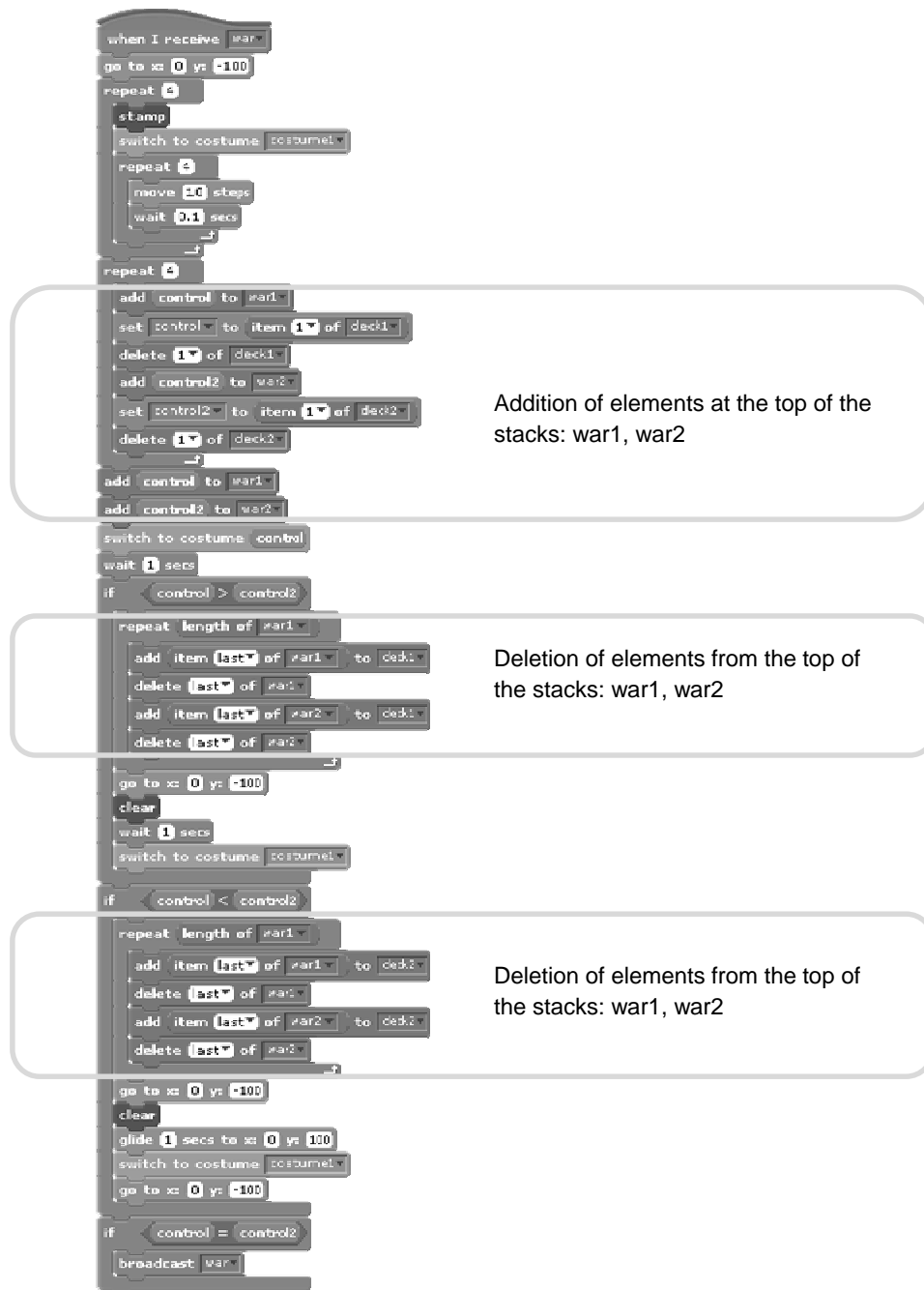


Figure 7. Card game “War” – the relation between last cards in stacks war1 and war2 determines the winner in the war

One of our students created a Scratch project with games and animation that presents data structures stack and queue, basic operations and data manipulation. In order to develop the game “Do the dishes”, which explains how data are processed in a stack, the programmer actually uses the structure stack. The stack is used to save the y coordinate of the dishes. (Figure 8) Every new plate appears on a new position that is higher than the top plate, i.e. the top of the stack saves the y coordinate of the last plate. When he player tried to take a plate that is not on the top of the pile, a message appears saying that they are going to break the dishes.[6]

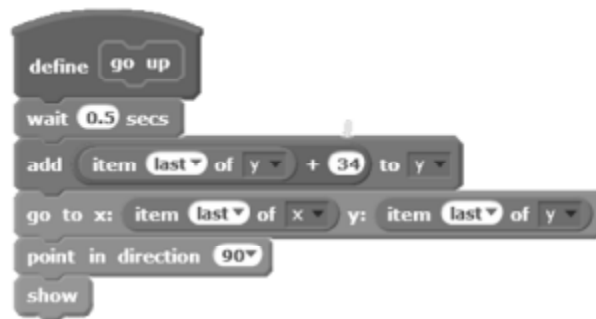


Figure 8. Project SteOp

As educators we want our learners to be able to formulate problems and to find “their solutions, represented in a form that can be effectively carried out by an information-processing agent.”[3] We know the theory, the code writing, but we should help young students to discover the benefits from a solid understanding of data structures and algorithms in a different way. We have a lot of interactive tools, we have our experience and we can provide the inspiration for the younger generation to find the place in the field of computer technologies and see a new wave of ambitious people, entering with new ideas and vision. We only need to play with them.

### References

1. Agrawal, R., Kurmaz, Z., Gudivada, V., Motivating Students to Learn Programming using Game Assignments, American Society for Engineering Education, 2012 [www.asee.org/public/conferences/8/papers/3083](http://www.asee.org/public/conferences/8/papers/3083)
2. California CEOs, educators and community leader to Governor Jerry Brown <https://code.org/action/brown-letter>
3. Cuny, J., Snyder, L. Demystifying computational thinking for non-computer scientists <http://cs.cmu.edu/~CompThink/resources/TheLinkWing.pdf>
4. <http://scratch.mit.edu/>

5. O'Neill, S., Howel, S., Scratch from scratch for schools and clubs, MIT, 2009
6. Staneva, G. SteOp – Scratch project, <http://scratch.mit.edu/projects/19375171/#player>

**Author:**

**Eleonora Stefanova Pavlova**, High School of Mathematics “D-r Petar Beron” – Varna, GSM: +359888925997, [eleonora.pavlova@gmail.com](mailto:eleonora.pavlova@gmail.com)

## DERIVATIVE OF FUNCTION WITH MAPLE

Nedka Pulova, Zlatka Mateva

**Abstract.** *In this paper we show how the computer algebra system Maple can be applied for teaching one of the basic concepts of calculus – the derivative of a function. The examples we present illustrate that calculating derivatives of functions with Maple goes hand-in-hand with the understanding of the mathematics lying behind. We use the Maple's graphical tools to explain the geometrical sense of the derivative, as well as, to illustrate the nonsmoothness of a function at the point, where the derivative does not exist. It is shown how the tutorial package "Calculus" can be invoked in the geometrical interpretation of the basic theorems of calculus – the Roll's theorem and the mean value theorem.*

**Keywords.** *Derivative and differential of function, Maple.*

### 1. Introduction

It is a negative tendency that in the last two decades the students avoid or simply abandon from studying mathematics. The knowledge and the skills of the students in mathematics dramatically decrease. On the other hand the computers are already accessible to the educational process. In order to make the mathematical objects more understandable and interesting some laboratory training exercises have been introduced. In that classes the mathematics is studied by the help of the contemporary computer algebra systems (CASs) Mathematica, Matlab, Maple etc.

Discussions on the advantages and disadvantages of one or other CAS can be found in the virtual Internet space. It is really true (to certain extent) if one has acquired some familiarity with one CAS to be easier to learn a new one. A comparative analysis of Matlab and Maple are given in [3], where on the base of one geometrical problem the basic features of these CASs are simultaneously revealed. The tutorial packages available in Maple and the capabilities of presenting the calculus in symbolic language make Maple an appropriate tool for studying mathematics.

The big computational potentials of the CASs can be used to free the mathematical classes from making long and tedious calculations that are traditionally made by hand setting in this way the possibility of pursuing new aims in studying and teaching mathematics. The main purpose is not only to solve a lot of problems but using certain examples to give the solutions and visualizations, which facilitates the process of understanding, so that the mathematical knowledge becomes clearer and deeper. In such a way the efforts can be much more focused on choosing the right mathematical questions and on interpreting results.

Teaching with CASs requires a new approach of making lessons (cf [1]). A workbook containing short instructions concerning the CAS opera-

tors and their possible applications with some solved problems involved would be very useful. Such a workbook saves the time, allows the students to work more effectively in class and as well as at home when for making their homework exercises.

There are textbooks of mathematics with extensive use of Maple (see e.g. [2, 4]). It is remarkable that the textbook [2] is entirely written in the Maple environment. The electronic version of this book provides a new possibility of studying mathematics in the way that the reader is allowed (directly from the textbook) to execute the Maple commands in order to solve the problems and make the exercises given in the textbook.

The aim of this paper is to give an idea of studying with Maple one of the basic concept of calculus – the derivative of a function. Led by the wish to get out of the plane of simply applying the Maple's *diff* operator we suggest seven problems, and then, we solve them by making use of the basic Maple's operators. We give also some auxiliary exercises and comments. For easy reading we give not only the commands applied but also the output they produce.

## 2. The derivative and the differential of a function with Maple

We suppose that the students know the definitions of the derivative and the differential of a function and the necessary and sufficient condition for the differentiability of a function, namely,  $f'_-(a) = f'_+(a)$ , where  $f'_-(a)$  and  $f'_+(a)$  are respectively the left-hand side and the right-hand side derivatives of the function  $f(x)$  at the point  $x = a$ .

### Problem 1:

**a)** Find the derivative of the function  $f(x) = \sqrt{x}$  at the point  $x = 1$  with Maple.

**b)** Find the differential of  $\sqrt{x}$  at  $x = 1$  for  $h = 0,2$ . Using the definition of the differential of a function at a point, find the approximate value of  $\sqrt{1,2}$ . Compare the result with the value of  $\sqrt{1,2}$  computed by the Maple's *sqrt*.

**c)** Find  $f'_+(0)$  and  $\lim_{x \rightarrow 0+} f'(x)$  and compare their values.

**a)** We give two ways of finding the derivative of a function:  
1. By the very definition (left column below).



2. By using successively the Maple's *diff* and *eval* operators. The operator *diff* gives the derivative  $f'(x)$ , the operator *eval* evaluates  $f'(0)$  (right column below).

$$\left[ \begin{array}{l} > f := x \rightarrow \sqrt{x}; \\ > \Delta f := x \rightarrow f(a+h) - f(a); \\ > a := 1; \\ > \text{Limit}\left(\frac{\Delta f(h)}{h}, h=0\right) = \lim_{h \rightarrow 0} \left(\frac{\Delta f(h)}{h}, h=0\right); \end{array} \right. \quad \begin{array}{l} f := x \rightarrow \sqrt{x} \\ \Delta f := x \rightarrow f(a+h) - f(a) \\ a := 1 \\ \lim_{h \rightarrow 0} \frac{\sqrt{1+h}-1}{h} = \frac{1}{2} \end{array} \quad \left[ \begin{array}{l} > \text{Diff}(f(x), x) = \text{diff}(f(x), x); \\ > \text{eval}(\%, x=1); \end{array} \right. \quad \begin{array}{l} \frac{d}{dx}(\sqrt{x}) = \frac{1}{2\sqrt{x}} \\ \frac{d}{dx}(\sqrt{x}) \Big|_{x=1} = \frac{1}{2} \end{array}$$

The students see that Maple uses the same notations of the derivative as in mathematics, namely,  $\frac{d}{dx}f$  for the derivative of  $f$  at an arbitrary point  $x$ , and  $\frac{df}{dx}\Big|_{x=a}$  for  $f'(a)$ .

**b)** Let us consider the solution in Maple.

According to the notation of the differential,  $df(a) = f'(a)h$ , we find the differential  $d\sqrt{x}\Big|_{x=1}$  for  $h=0,2$ . Note that the latter operator is not an assignment operator.

$$\left[ \begin{array}{l} > f := x \rightarrow \sqrt{x}; \\ > h := 0.2; \\ > 'd\sqrt{x}\Big|_{x=1}' = h.\text{subs}(x=1, \text{diff}(f(x), x)); \end{array} \right. \quad \begin{array}{l} f := x \rightarrow \sqrt{x} \\ h := 0.2 \\ d\sqrt{x}\Big|_{x=1} = 0.1000000000 \end{array}$$

By making use of the formula  $f(a+h) \approx f(a) + df(a)h$ , for  $a=1$  and  $h=0,2$ , we obtain  $\sqrt{1,2} \approx 1,1$ . The approximate value calculated with the Maple's *sqrt* is  $\sqrt{1,2} \approx 1,095445115$ .

```

> a := 1;
> f(a) + h.subs(x = 1, diff(f(x), x));
> sqrt(1.2);

```

$a := 1$   
 $1.100000000$   
 $1.095445115$

c) The right-hand side derivative of  $\sqrt{x}$  at  $x = 0$  is found by applying the *limit* operator with a *right* option.

```

> restart :
> f := x -> sqrt(x);
> Delta f := x -> f(a + h) - f(a);
> a := 0;
> Limit(Delta f(h)/h, h = 0, right) = limit(diff(f(x), x), x = 0, right);

```

$f := x \rightarrow \sqrt{x}$   
 $\Delta f := x \rightarrow f(a + h) - f(a)$   
 $a := 0$   
 $\lim_{h \rightarrow 0^+} \frac{1}{\sqrt{h}} = \infty$

The *limit* operator finds the right-hand side limit of the derivative,  $\lim_{x \rightarrow 0^+} f'(x)$ .

```

> Limit(Diff(f(x), x), x = 0, right) = limit(diff(f(x), x), x = 0, right);

```

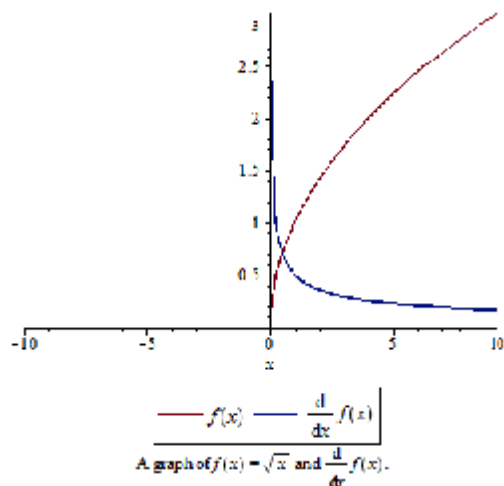
$\lim_{x \rightarrow 0^+} \left( \frac{d}{dx} \sqrt{x} \right) = \infty$

**Note.** It follows from c) that the right derivative at  $x = 0$  and the right limit of the derivative when  $x \rightarrow 0$  are equal, i.e.,  $\lim_{x \rightarrow 0^+} f'(x) = f'_+(0)$ . We will use the equalities  $\lim_{x \rightarrow x_0^-} f'(x) = f'_-(x_0)$  and  $\lim_{x \rightarrow x_0^+} f'(x) = f'_+(x_0)$  for finding the derivatives at a point and will explain these equalities in more details in section 5.

There is a special command, *Derivative(Plot)*, of the package *Student[Calculus1]*, that produces a plot combining the graphs of the func-

tion and its derivative. This plot can help in clarifying the notion of the derivative as a function of the independent variable.

```
> with(Student[CalculusI]):
> DerivativePlot(sqrt(x));
```



The next problem as well as the point b) of the Problem 1 (see also the exercise below) could be useful for better understanding the definition of the differential.

**Problem 2.** Using the definition of the differential  $df$  of a function  $f(x)$  as a linear part of the function's change  $\Delta f$ , find the differential of the function  $f(x) = x^3 - 2x + 7$  at an arbitrary point  $x = a$  for an arbitrary change of the independent variable  $\Delta x = h$ .

The operator *collect* views the change  $\Delta f$  of the function as a polynomial of  $h$  and the operator *coeff* extracts the coefficient of  $h$  in the polynomial.

```
> f := x -> x^3 - 2x + 7;
> Δf := x -> f(a + h) - f(a);
> Δf := sort(collect(Δf, h), h, ascending);
> 'df|x=a' = coeff(Δf, h);
```

$$f := x \rightarrow x^3 - 2x + 7$$

$$\Delta f := x \rightarrow f(a + h) - f(a)$$

$$\Delta f := (3a^2 - 2)h + 3ah^2 + h^3$$

$$df|_{x=a} = (3a^2 - 2)h$$

It must be pointed out that the solution obtained works only in the case when the function  $f(x)$  is a polynomial. It would be useful to compare the above result with the next one.

$$\left[ \begin{array}{l} > 'df|_{x=a}' = \text{subs}(x = a, \text{diff}(f(x), x))h; \\ & \qquad \qquad \qquad df|_{x=a} = (3a^2 - 2)h \end{array} \right.$$

**Exercise.** Find the approximate value of  $\cos 61^\circ$  using the definition of the differential of a function. Compare the result with the value of  $\cos 61^\circ$  computed by the Maple's `cos`.

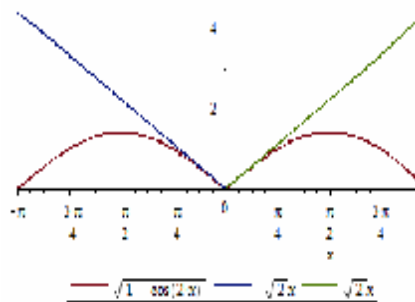
Having in mind the way of finding derivatives by applying in sequence `diff` and `eval` (as in Problem 1 a)) one can be easily misled that any problem can be solved in this way. The next example convinces us in the need of using the definition of the derivative.

**Problem 3.** Find the derivative of the function  $\sqrt{1 - \cos 2x}$ .

The operator `diff` produces  $f'(x) = \frac{\sin 2x}{\sqrt{1 - \cos 2x}}$ . Unfortunately, as

the denominator vanishes at  $x = 0$ , this expression can not be applied for calculating the derivative at this point. Now we turn directly to the definition of the derivative.

$$\left[ \begin{array}{l} > f := x \rightarrow \sqrt{1 - \cos(2x)}; \\ & \qquad \qquad \qquad f := x \rightarrow \sqrt{1 - \cos(2x)} \\ > \text{Diff}(f(x), x) = \text{diff}(f(x), x); \\ & \qquad \qquad \frac{d}{dx}(\sqrt{1 - \cos(2x)}) = \frac{\sin(2x)}{\sqrt{1 - \cos(2x)}} \\ \Delta f := x \rightarrow f(a + h) - f(a); \\ & \qquad \qquad \Delta f := x \rightarrow f(a + h) - f(a) \\ > a := 0; \\ & \qquad \qquad a := 0 \\ > \text{Limit}\left(\frac{\Delta f(h)}{h}, h = 0, \text{left}\right) = \lim_{h \rightarrow 0^-} \left(\frac{\Delta f(h)}{h}, h = 0, \text{left}\right); \\ & \qquad \qquad \lim_{h \rightarrow 0^-} \frac{\sqrt{1 - \cos(2h)}}{h} = -\sqrt{2} \\ > \text{Limit}\left(\frac{\Delta f(h)}{h}, h = 0, \text{right}\right) = \lim_{h \rightarrow 0^+} \left(\frac{\Delta f(h)}{h}, h = 0, \text{right}\right); \\ & \qquad \qquad \lim_{h \rightarrow 0^+} \frac{\sqrt{1 - \cos(2h)}}{h} = \sqrt{2} \end{array} \right.$$



From the definition of the one-side derivative at  $x=0$ , we find  $f'_-(0) = -\sqrt{2}$  and  $f'_+(0) = \sqrt{2}$ , implying that the function  $\sqrt{1-\cos 2x}$  is not differentiable at  $x=0$ . In the next section the differentiability of a function is considered from geometrical point of view.

### 3. The geometrical sense of the derivative, differentiability and nondifferentiability of a function and Maple

It is an undeniable fact that the computer visualization goes hand-in-hand with the understanding of the mathematical ideas. Graphical potentials of Maple makes it possible to observe how the tangent of the function's graph at the point  $(a, f(a))$  is obtained as a limit of a sequence of secant lines. Maple offers the following operators:

*FunctionSlopePlot* *FunctionSlopeTutor()* of the package *Student[Precalculus]*  
*TangentSecantTutor()* *TangentTutor()* of the package *Student[Calculus1]*.

The operators of the package *Precalculus* illustrate the convergence of the secant lines passing through the points  $(a, f(a))$  and  $(a+h, f(a+h))$ , first for  $h \rightarrow 0^+$  and then for  $h \rightarrow 0^-$ , while the operators of the package *Calculus* give the secant line only in the case when  $h \rightarrow 0^+$ .

The geometrical sense of the derivative as the slope of the tangent line to the graph of the function is illustrated by the next problem.

**Problem 4.** Find the derivative of the function  $f(x) = \begin{cases} x^2 + x + 1, & x \leq 0 \\ x^3 + x + 1, & x > 0 \end{cases}$ . Find  $f'(0)$  and the slope of the tangent line to the graph of the function at the point  $(0, f(0))$ .

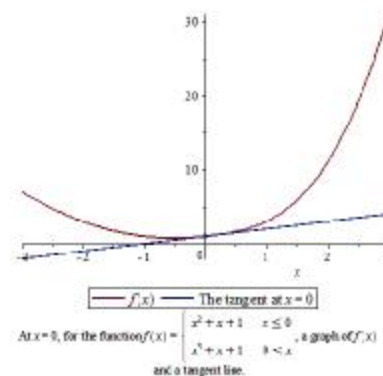
In Maple the given function is defined with *piecewise* operator. From the result obtained by the operator *diff* it follows directly that the derivative of the function at  $x=0$  exists and its value is  $f'(0) = 1$ . Despite this it is useful to ensure that the sufficient condition  $f'_-(0) = f'_+(0)$  at  $x=0$  is fulfilled. The operator *limit* finds the one-side limits of the derivative,  $f'_-(0) = \lim_{x \rightarrow 0^-} (2x+1) = 1$  and  $f'_+(0) = \lim_{x \rightarrow 0^+} (3x^2+1) = 1$ , and hence, it follows that the derivative at  $x=0$  exists and  $f'(0) = 1$ .

For calculating the slope of the tangent line we activate the package *geomtry*. We use in succession: (1) the operator *Tangent* for finding the equation of the tangent line to the graph at the point  $(0, f(0))$ ; (2) the operator *line* for defining a line  $l$  that is the tangent to the graph at the point  $(0, f(0))$ , and in the end, (3) the operator *slope* that gives the slope of the line  $l$ . Thus we have the slope of the tangent line. In this way the students can see also that the derivative  $f'(0)$  is equal to the slope of the tangent line at  $(0, f(0))$ .

The operator *tangent* executed with the option *plot* gives the graph of the function and the tangent line on the same plot.

```
> F := piecewise(x ≤ 0, x2 + x + 1, x > 0, x3 + x + 1)
                                F := { x2 + x + 1  x ≤ 0
                                { x3 + x + 1  x > 0
> diff('F', x) = diff(F, x)
                                ∂
                                ∂x F = { 2x + 1  x ≤ 0
                                { 3x2 + 1  x > 0
> limit(diff(F, x), x = 0, right)
                                1
> limit(diff(F, x), x = 0, left);
                                1
> 'k' = eval(diff(F, x), x = 0);
                                k = 1
```

```
-----
> with(geomtry): with(Student[Calculus1]):
> Tangent(F, x = 0);
                                x + 1
> line(l, y = Tangent(F, x = 0), [x, y]);
                                l
> 'l' = Equation(l);
                                l = (y - x - 1 = 0)
> 'k' = slope(l);
                                k = 1
> Tangent(F, x = 0, output = plot, -3..3);
```



**Note.** The operator  $Tangent(f(x), x = 0, output = slope)$ , applied to the function  $\sqrt{1 - \cos 2x}$ , gives that the slope at the point  $x = 0$  is not defined. Really, we have proved (see problem 3) that the function  $\sqrt{1 - \cos 2x}$  is not differentiable at  $x = 0$ , and hence, the tangent line

does not exist. The lines  $y = -\sqrt{2}x$  and  $y = \sqrt{2}x$  are respectively the left and the right tangent lines at  $(0, f(0))$ .

**Exercise.** Find the derivative of the functions:

$$1) f(x) = \begin{cases} x, & x \leq 0 \\ \ln(x+1), & x > 0 \end{cases}$$

$$2) f(x) = |x| \quad 3) f(x) = \sqrt[3]{x^2 - 1} \quad 4) f(x) = \begin{cases} 1-x & x < 0 \\ e^x, & x \geq 0 \end{cases}$$

$$5) f(x) = \begin{cases} x^2 e^{-x^2}, & x < 1 \\ \frac{1}{e}, & x \geq 1 \end{cases}$$

#### 4. The differentiation rules with Maple

For finding derivatives, Maple offers the command *Rule* with the options: *sum*, *constantmultiple*, etc. For the sake of completeness we give examples of using different options of this operator. By the authors' knowledge the command *Rule* is not among the commands the students like to use. However, finding derivatives with the *Rule* operator is a very helpful exercise for working with compositions of functions and for applying the differentiation rules.

**Problem 5.** Find the derivatives: **a)**  $(3 \sin x)'$  **b)**  $(\sin 3x)'$

**a)** We use in succession two options of *Rule*: *constantmultiple* and *sin*

```
> with(Student[Calculus1]):
> F := x -> 3 sin(x);
   F := x -> 3 sin(x)
> Rule[constantmultiple](Diff(F(x), x));
      d
      dx
      (3 sin(x)) = 3
      d
      dx
      (cos(x))
> Rule[sin](%);
      d
      dx
      (3 sin(x)) = 3 cos(x)
```

**b)** Considering this function as a composition of two functions it can be differentiated using successively the *chain*, *sum* and *identity* options of the operator *Rule*.

$$\begin{array}{l}
 > F := x \rightarrow \sin(3x); \\
 & \quad F := x \rightarrow \sin(3x) \\
 > \text{Rule}[\text{chain}](\text{Diff}(F(x), x)); \\
 & \quad \frac{d}{dx}(\sin(3x)) = \left( \frac{d}{dX} \sin(X) \right) \bigg|_{X=3x} \left( \frac{d}{dx}(3x) \right) \\
 > \text{Rule}[\sin](\%); \\
 & \quad \frac{d}{dx}(\sin(3x)) = \cos(3x) \left( \frac{d}{dx}(3x) \right) \\
 > \text{Rule}[\text{constantmultiple}](\%); \\
 & \quad \frac{d}{dx} \sin(3x) = 3 \cos(3x) \left( \frac{d}{dx}(x) \right) \\
 > \text{Rule}[\text{identity}](\%); \\
 & \quad \frac{d}{dx} \sin(3x) = 3 \cos(3x)
 \end{array}$$

Another operator, *DiffTutor*, from the package *Student[Calculus1]*, with the same consequence of options *chain*, *sum* and *identity*, provides a possibility for solving the Problem 5 in interactive regime.

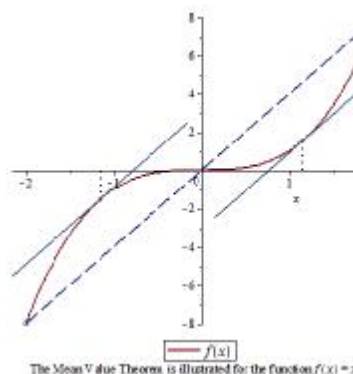
**Exercise.** Find  $(x^{\sin x})'$  by using the required options: *rewrite*, *chain*, *exp*, etc.

## 5. The basic theorems of calculus and their applications with Maple

The mean value theorem and the Rolle's theorem can be studied with Maple. The operator *MeanValueTheorem* plots the graph of a function  $f(x)$ ,  $x \in [a, b]$ , and the tangent lines to that graph, which slope is equal to  $\frac{f(b) - f(a)}{b - a}$ . On using the *option=points*, this operator gives the points  $x$  from the segment  $(a, b)$  (ensured by the mean value theorem), for which  $\frac{f(b) - f(a)}{b - a} = f'(x)$ . The example below is for the function  $f(x) = x^3$  on the interval  $x \in [-2, 2]$ .



```
> with(Student[Calculus1]):
> MeanValueTheorem(x^3, -2..2);
> MeanValueTheorem(x^3, -2..2, output = points);
```

$$\left[-\frac{2}{3}\sqrt{3}, \frac{2}{3}\sqrt{3}\right]$$


A helpful consequence of the mean value theorem, that we have already used for solving the Problem 4, is the next Proposition (cf [6]).

**Proposition.** Let the function  $f(x)$  is differentiable in the interval  $(a, b)$  except (eventually) the point  $x_0 \in (a, b)$  and continuous at  $x_0$ . Then  $f'_-(x_0)$  exists and  $f'_-(x_0) = \lim_{x \rightarrow x_0^-} f'(x)$  if  $\lim_{x \rightarrow x_0^-} f'(x)$  exists (finite or infinite). Similarly  $f'_+(x_0) = \lim_{x \rightarrow x_0^+} f'(x)$ .

Let us return to the Problem 3. We saw that the function  $\sqrt{1 - \cos 2x}$  has no derivative at the point  $x = 0$ . At the same time the denominator of the derivative becomes zero at  $x = 0$ . On this observation the students can make (and they really do) the conclusion that this is a general rule. The next problem shows that this is actually not true (see e.g. [5]).

**Problem 6.** Find the derivative of the function  $f(x) = x\sqrt{1 - x^2} \sin x^2$ ,  $x \in (-\sqrt{3}, \sqrt{3})$ . Prove the differentiability of this function at  $x = 0$  and its not differentiability at  $x = 1$ .

The expression of the derivative, given by the operator *diff*, is defined for  $x \neq 0, 1$ . On finding the left and the right limits of the derivative at the points  $x = 1$  and  $x = 0$  we obtain  $f'(0) = 0$  while  $f'(1)$  does not exist. Geometrically this means that the tangent to the graph at the point

$(1,0)$  coincides with the coordinate axis  $Ox$ , while the tangent at the point  $(0,0)$  does not exist (see Fig. 1).

```

> f := x -> x*sqrt(1-x)^2*sin(x^2);
      f := x -> x*sqrt(1-x)^2*sin(x^2)
> Diff(f, x) = diff(f(x), x);
      ∂
      ∂x f = sqrt(1-x)^2*sin(x^2) + 1/2 * x * (-2(1-x)*sin(x^2) + 2(1-x)^2*cos(x^2)*x) / sqrt(1-x)^2*sin(x^2)
> limit(diff(f(x), x), x = 0);
      0
> limit(diff(f(x), x), x = 1, left);
      -sqrt(sin(1))
> limit(diff(f(x), x), x = 1, right);
      sqrt(sin(1))
> plot(f(x), x = -2..2);

```

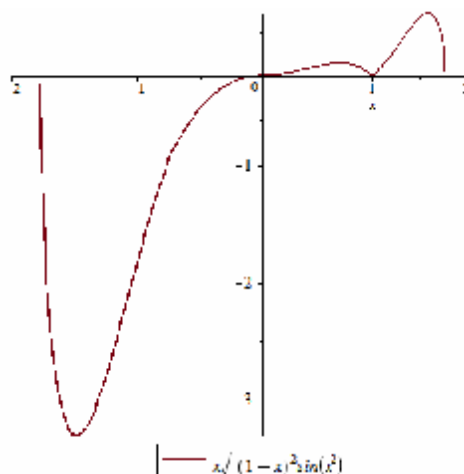


Fig. 1

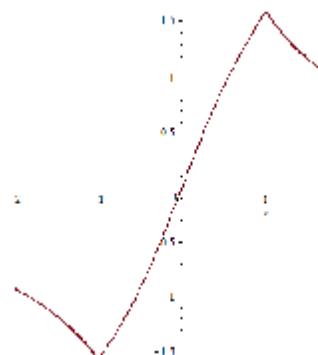
We recommend solving the next problem first by hand or in parallel with Maple.

**Problem 7.** Find the left and the right derivatives of the function  $\arcsin \frac{2x}{1+x^2}$  at the points  $x = -1$  and  $x = 1$ . Prove that the function is not differentiable at  $x = -1$  and  $x = 1$ .

The result below is obtained by the application of *diff* and *simplify*.

$$\left[ \begin{array}{l} > f := x \rightarrow \arcsin\left(\frac{2x}{1+x^2}\right); \\ & f := x \rightarrow \arcsin\left(\frac{2x}{1+x^2}\right) \\ > \text{simplify}(\text{diff}(f(x), x)); \\ & -\frac{2(x^2-1)}{(x^2+1)^2 \sqrt{\frac{(x^2-1)^2}{(x^2+1)^2}}} \end{array} \right.$$

The expression calculated by the *simplify* operator can be in fact further simplified. It is a frequent mistake to cancel the root and the exponent, but the experienced students know that there is a serious reason preventing Maple from making such a reduction. Actually Maple needs some additional information that can be supplied by calling the optional operator *assuming*.

$$\left[ \begin{array}{l} > \text{simplify}(\text{diff}(f(x), x) \text{ assuming } x \geq 1; \\ & -\frac{2}{x^2+1} \\ > \text{simplify}(\text{diff}(f(x), x) \text{ assuming } -1 < x < 1; \\ & \frac{2}{x^2+1} \\ > \text{simplify}(\text{diff}(f(x), x) \text{ assuming } x \leq -1; \\ & -\frac{2}{x^2+1} \\ > \text{plot}(f(x), x = -2..2); \end{array} \right.$$


It follows from the Maple's output and the Proposition that the considered function has not derivatives at  $x = -1$  and  $x = 1$ , which is in full correspondence with the nonsmoothness of the graph at these points.

The next problem shows the importance of providing all the conditions required by the theorems involved.

**Problem 8.** Apply the operator *MeanValueTheorem* to the function  $f = |x|$  in the interval  $[a, b]$  where  $a < 0 < b$ . Prove that the equation  $\frac{f(b) - f(a)}{b - a} = f'(x)$  has not solution for  $x \in (a, b)$ . Explain this negative result.

**Problem 9.** Apply the operator *diff* to the function

$$f(x) = \begin{cases} x \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0 \end{cases}$$

Explain the output obtained.

## 6. Conclusion

In this paper we give an idea of studying the derivative of a function with Maple. We have shown, by solving several problems, how Maple can be used to facilitate the understanding of the notion of the derivative and its geometrical meaning, as well as, the notion of a smooth and nonsmooth function at a point. It is shown how Maple can be used in the process of studying the differentiation rules, etc.

There are some technical difficulties arising in studying with Maple. It is important to be very careful to avoid syntax mistakes that take a lot of time to be corrected. On the other hand the frequent use of the help menu deflects the students' attention from the main stream of the lesson.

It is advisory for students that whenever it is possible to find the solution of the problem by hand using in parallel the Maple's commands. It would be helpful that the students write short Maple procedures, as for example, such a procedure on the mean value theorem would lead the students to the idea of numerical solving of equations.

Studying mathematics in the lab classes with Maple supposes intensive activity of the students that corresponds to the didactical principles of teaching and learning.

## References

- [1] S. Karakoleva, E. Veleva. Practical course on numerical methods for engineering students. Proc. of the 43-rd Spring Conference of the Union of Bulgarian Mathematicians, Borovetz, April, 2-6, 2014, 260-265.
- [2] K. Rosen. Exploring discrete mathematics using Maple. [http://highered.mcgraw-hill.com/sites/0073383090/student\\_view0/exploring\\_discrete\\_mathematics\\_using\\_maple.html](http://highered.mcgraw-hill.com/sites/0073383090/student_view0/exploring_discrete_mathematics_using_maple.html).
- [3] Z. Mateva. Applying the Maple and Matlab systems in teaching the topic "Equations for straight lines and planes in the space". Proc. of the 43-rd Spring Conference of the Union of Bulgarian Mathematicians, Borovetz, April, 2-6, 2014, 286-291.
- [4] J. Oprea. Differential geometry and its applications. Prentice-Hall, New Jersey, 1997.
- [5] И. Виноградова, С. Олехник, В. Садовничий. Задачи и упражнения по математическому анализу. Изд. Моск. унив., 1988.
- [6] А. Тер-Крикоров, М. Шабунин. Курс математического анализа. М. Наука, 1988.

**Authors:**

**Nedka Pulova**, PhD, Department of Mathematics and Foreign Languages, Section of Mathematics, Technical University of Varna, Studentska Str. 1, 9010, Varna, Bulgaria, tel. 052 383 250, e-mail: [vpulov@yahoo.com](mailto:vpulov@yahoo.com)

**Zlatka Mateva**, PhD, Department of Mathematics and Foreign Languages, Section of Mathematics, Technical University of Varna, Studentska Str. 1, 9010, Varna, Bulgaria, 052 383 228, e-mail: [ziz@abv.bg](mailto:ziz@abv.bg)

## MOODLE GAMIFICATION

Milen Sotirov

**Abstract:** *Gamification is using game mechanics to engage, motivate action, promote learning, and etc. When gamification mechanics are applied in a Moodle course can engage more students and can provide opportunities for fun and interactive learning.*

**Keywords:** *moodle, gamification, game mechanics, game dynamics, eLearning, badges*

### What is Gamification

Gamification is using game-based mechanics and game-thinking to engage people, motivate action, promote learning, and solve problems [1]. When gamification mechanics are applied for educational purposes in a Moodle course, the teachers can increase the student's engagement and they can provide opportunities for learning in interactive and fun environment [2]. Gamification as a term originated in the digital media industry. The first documented uses dates back to 2008, but gamification only entered widespread adoption in the second half of 2010, when several industry players and conferences popularized it. It is also—still—a heavily contested term; even its entry into Wikipedia has been contested. Within the video game and digital media industry, discontent with some interpretations have already led designers to coin different terms for their own practice (e.g., gameful design) to distance themselves from recent negative connotations[3]. Until now, there has been hardly any academic attempt at a definition of gamification. Current uses of the word seem to fluctuate between two major ideas. The first is the increasing societal adoption and institutionalization of video games and the influence games and game elements have in shaping our everyday life and interactions. Game designer Jesse Schell summarized this as the trend towards a Gamepocalypse, "when every second of your life you're actually playing a game in some way"[4]. As a result of a series of changes in demographics, technology, and the competitive landscape, smart companies, as well as nonprofits and governments, are increasingly turning to play and games as a way of radically reinventing their organizations. They are engaging customers as never before, aligning employees, and driving innovation that seemed virtually impossible only a decade ago. They've realized that their strength is in marshaling the intelligence, motivation, and—most critically—engagement of their communities to drive their business objectives [5].

### Game Mechanics

Game mechanics are constructs of rules intended to produce a game or gameplay. All games use mechanics; however, theories and styles differ as to their ultimate importance to the game. In general, the

process and study of game design are efforts to come up with game mechanics that allow for students playing a game to have an engaging, but not necessarily fun, experience [2]. The interaction of various game mechanics in a game determines the complexity and level of student's interaction in the game, and in conjunction with the game's environment and resources determine game balance. Some forms of game mechanics have been used in games for centuries, while others are relatively new, having been invented within the past decade. Game Mechanics & Game dynamics are closely related and sometimes used interchangeably. Game mechanics are the various actions, behaviors, and control mechanisms that are used to “gamify” an activity – the aspects that, taken together, create a compelling, engaging student experience. The compelling, motivational nature of this experience is, in turn, the result of desires and motivations that's called game dynamics [6] [7].

Table 1. Game mechanics and Game dynamics

Game mechanics include:	Game dynamics include:
Points	Reward
Levels	Status
Challenges	Achievement
Virtual goods and spaces	Self-expression
Leaderboards	Competition
Gifts and charity	Altruism



Figure 1. Elements of gamification in a Moodle course [8]

	Reward	Status	Achievement	Self Expression	Competition	Altruism
Points	●	●	●		●	●
Levels		●	●		●	
Challenges	●	●	●	●	●	●
Virtual Goods	●	●	●	●	●	
Leaderboards		●	●		●	●
Gifting & Charity		●	●		●	●

Figure 2. Lecturers can choose what to integrate in their course [9]

### The Best Game Element for Moodle Course

#### • Badges

Badges are a good way of celebrating achievement and showing progress. Badges can be awarded based on a variety of chosen criteria and are fully compatible with Mozilla Open Badges. Badges created in Moodle can be displayed on a user's profile or pushed to their Open Badges backpack and public badge collections in their backpack and in their Moodle profile [10]. Earning rewards by successfully completing achievements set by teachers is a valuable way of showing what the students know and taking it with them where they go. Students can see which rewards they have earned and what is required to receive additional rewards, providing insight into learning progression toward defined competencies. Teachers can easily see which students have reached learning milestones [11]. They provide positive reinforcement for the targeted behavior.

There are two categories of badges [10]:

§ **Site badges** - available to students site-wide and related to the site wide activities, like finishing a set of courses.

§ **Course badges** - available to students enrolled in the course and related to the activities that happen inside the course.

#### • Badgeville

The Behavior Platform empowers Universities to build more personalized, rewarding experiences for the students and employees on top of your websites, enterprise applications, social networks, and communities. Named as the best platform as a service in 2013 by the Software & Information Industry Association (SIIA), The Behavior Platform helps to create a comprehensive program for user engagement that increases customer loyalty and enhances employee performance.[12]



- **Mozilla Open Badges**

The Mozilla Open Badges (Open Badge Infrastructure or OBI) project is a program by Mozilla that issues digital badges to recognize skills and achievements. The badge structure allows one to display real-world achievements and skills which may help with future career and education opportunities. The OBI allows badge issuers and developers to build badges, and allows online learners to choose from a number of different pathways for development [13].



Figure 3. Earned badges

- **Mozilla Backpack**

The students can earn badges from anywhere, and then they can share them wherever they want — on social networking profiles, job sites or website. Mozilla Backpack makes it easy to earn badges from multiple sources, both online and offline. Then the student can sort them into categories and to choose where they're shared [14][15].

- **Progress Bar or XP Bar**

An important part of providing feedback to students in games or eLearning is to let them know how much progress they've made. There are many ways to represent this, but the most effective are always represented graphically [10].

The Progress Bar is a time-management tool for students. It visually shows what activities/resources a student is supposed to interact with in a course. It is color coded so students can quickly see what they have and have not completed/viewed. The teacher selects which pre-existing activities/resources are to be included in the Progress Bar and when they should be completed/viewed. Ordering can be done by times/deadlines or by the ordering of activities in the course. There is an overview page allowing teachers to see the progress of all students in a class, which is helpful for finding students at risk [10].

- **Checklist**

The **Checklist Module** is an add-on that allows a teacher to create a checklist for their students to work through. The students can tick off each item on the list, as they complete it and the resulting progress of the class can be monitored by the teacher. The student is also rewarded with a progress bar that shows how far through the checklist they have managed to get. [10]

- **Ranking**

Leaderboards bring in the social aspect of points and badges, by displaying the students on a list typically ranked in descending order with the greatest number of points at the top. It creates a sense of competition among the students. The ranking block works together with the Moodle course completion. The ranking block monitors activities and add points to the students based on accomplishments. [10]

- **Levels**

To enroll in the course students must take placement test to determine their level of experience [16], based on the outcome they are placed in course for advanced or beginner.

- **Relationships**

Relationships are game mechanics based on the motivational driver of connection. We are social beings, and relationships have a strong effect on how we feel and what we do [17].

- **Achievements, Goals, Quests, Missions and Challenges (with epic meaning)**

They can give students a way to brag about what they've done indirectly as well as add challenge and character to a game. They are often considered as "locked" until the students met the series of tasks that are required to "unlock" [12].

- **Challenge** is a powerful game mechanic to motivate students to action, especially if they believe they are working to achieve something great, something awe-inspiring, and something bigger than themselves. A structured way to accomplish something [7].

- **Quest** unlocks helpful hidden information.

- **Hidden info** is an intentionally hidden message, joke, hint or feature within a game. In a Moodle course, this can contain an interesting video, a new piece of information to make a task easier, or an extra quest (i.e., resource or activity) that is not available until that point [18].

- **Journey**

The journey game mechanic recognizes that the student is on a personal journey and incorporates this element into the experience [19].

- **Constraints (with Urgent Optimism)**

It has been shown that constraints such as deadlines, when combined with urgent optimism, motivate students to action. Urgent Optimism refers to extreme self-motivation. It is the desire to act immediately to tackle an obstacle combined with the belief that we have a reasonable hope of success [19].

- **Moodle Conditional Access - Content-unlocking**

Conditional activities enable teachers to restrict the availability of any activity or even a course section according to certain conditions such as dates, grade obtained, or activity completion. When it is enabled by the administrator, a "Restrict access" setting appears in the settings of activities or resources and also when editing a course section [10].

- **Moodle Completion Tracking**

In order to use the course completion features, it has to be enabled in the site administration settings before the course completion link will appear in the course administration menu [10].

- **Game book**

A game book allows the reader to participate in the story by making effective choices. The narrative branches along various paths through the use of numbered paragraphs or pages [20].

- **Chat**

Student response system that empowers teachers to engage their classrooms through a series of educational exercises and games via smartphones, laptops, and tablets [21].

- **Embedded class room system - <http://www.classdojo.com/>**

ClassDojo is a classroom tool that helps teachers improve behavior in their classrooms quickly and easily. It also captures and generates data on behavior that teachers can share with parents and administrators. Better learning behaviors, smoother lessons and hassle-free data - and it's free! [22]

- **Embedded - <https://habitrrpg.com/>**

HabitRPG is an online time management application. The game is designed as a self-improvement web application to help the player (student) keep track of and remain motivated by overlaying game mechanics onto a player's curriculum goals. The game is laid out in the form of a RPG, in which the student collects items such as gold and armor to become more powerful. Rewards are achieved through maintaining real-life goals, in the form of Habits, Dailies and To-Do's. [23]

- **Embedded – <http://chorewars.com/>**

Whenever one of the students in the class completes a task, he can log in and claim the XP [24].

### • Additional Rewards

The good learning experience, knowledge and the grade given to students are the most important rewards but every student wants more. The Additional Rewards are the first thing they will look up.



Figure 4. Example for student's rewards

### Conclusion

The Benefits of Gamification are many and varies from Industry to Industry. Game Mechanics and Game Features are designed to boost certain Metrics. In Gamification it is critical to monitor the performance of your Metrics closely to insure you are getting the best results possible and that the students are happy. When gamification is used in Moodle courses, teachers discover new ways to teach students by using game mechanics that are familiar to them like: Collaboration, Safe Failure, Engagement, Student choice and Alternative rewards.

Gamification helps:



Figure 5. Growth Statistic [25]

### Resources

[1] Kapp, Karl. (May 1, 2012) The Gamification of Learning and Instruction: Game-Based Methods and Strategies for Training and Education. San Francisco: Pfeiffer. ISBN: 978-1-118-09634-5. 336 pages.

[2] <http://badgeville.com>

- [3] McGonigal, J. We don't need no stinkin' badges: How to re-invent reality without gamification. Presentation at GDC 2011.
- [4] Schell, J. Visions of the Gamepocalypse. Presentation, Long Now Foundation, San Francisco, CA, July 27, 2010.
- [5] Zichermann Gabe, 2013, The Gamification Revolution: How Leaders Leverage Game Mechanics to Crush the Competition, New York : McGraw-Hill Education.
- [6] Friendjsr [friendjsr.blogspot.com](http://friendjsr.blogspot.com)
- [7] Bunchball inc. <http://www.bunchball.com/>
- [8] Yu-kai Chou [www.yukaichou.com](http://www.yukaichou.com)
- [9] Intouch Solutions <http://www.intouchsol.com/blog/bringing-the-power-of-gamification-to-pharma>
- [10] Moodle™ <https://moodle.org/>
- [11] Northern Illinois University <http://www.niu.edu/blackboard/upgrade/>
- [12] Badgeville <http://www.badgeville.com/>
- [13] Mozilla.org Mozilla Open Badges
- [14] Mozilla.org <http://backpack.openbadges.org/>
- [15] Mozilla.org <http://www.openbadges.org/>
- [16] Prince George's Community College <https://www.pgcc.edu>
- [17] UX Magazine <http://uxmag.com/articles/five-steps-to-enterprise-gamification>
- [18] Gamepedia [gamepedia.com](http://gamepedia.com)
- [19] UX Magazin [uxmag.com](http://uxmag.com)
- [20] Learn Out Live [learnoutlive.com](http://learnoutlive.com)
- [21] Socrative [www.socrative.com/](http://www.socrative.com/)
- [22] ClassTwist, Inc. [www.classdojo.com](http://www.classdojo.com)
- [23] HabitRPG <https://habitrpg.com>
- [24] Chore Wars <http://chorewars.com/>
- [25] Centro de Innovación BBVA <https://www.centrodeinnovacionbbva.com/en/magazines/innovation-edge/publications/5-gamification/posts/34-gamification-gaming-gamification>

**Author:**

**Milen Sotirov**, Expert SEO/SEM, Institute of Technology, VFU  
 Phone: +359-52-359-537  
 Email: [msotirov@vfu.bg](mailto:msotirov@vfu.bg)

## ADAPTABILITY TO THE KNOWLEDGE OF THE STUDENTS BY USING OF SCORM E-CONTENT

Asya Stoyanova, Todorka Glushkova

**Abstract.** *The article presents some approaches to adaptation of e-learning environment by using the SCORM -based training materials. We will review the trends in the SCORM standard and will share our experience in the design and creation of SCORM- based learning resources, as well as their application in the training process of students of bachelor and master programs in the Faculty of Mathematics, Computer Science and Information Technology of Plovdiv University.*

**Keywords:** SCORM, adaptability, DeLC, eLearning

### INTRODUCTION

The information society requires the application of new methods and approaches to education of students [4]. The approach that we follow is related to the study of adaptability as a key feature of any e-Learning system. Adaptability to students' knowledge of each e-learning system is a main indicator of its effectiveness [1], [2]. It can be seen from different viewpoints, but we will concentrate our attention mainly on those aspects which stem from the pedagogical practice and experience. The aims of this article are to present the general model of adaptive education through SCORM<sup>1</sup>-based courses and to share experience of the authors in the implementation of these e-learning resources in the portal for electronic and distance training DeLC [5].

**Adaptability** is an abstract concept which can have different specific forms of manifestation in e-Learning systems. There are different definitions of this concept according to the specific characteristics and goals of any such system. We will consider the adaptability as feature of the system that ensures maximum satisfaction of students and teachers in the e-learning process. The main elements of the adaptive model are "condition-action" rules that change the parameters of the environment and realize the adaptation to a user's knowledge, goals, abilities, preferences, etc. The implementation of the model requires the consideration of various aspects of adaptability of horizontal and vertical principles. The first one is connected with adaptation to: students' knowledge; students' goals and plans; their cognitive characteristics and emotional types; temporal characteristics of the training process; mode of access to learning resources etc. The second one is based on the classification of the species' adaptability to the level of implementation

---

<sup>1</sup> SCORM-Sharable Content Object Reference Model

and realization in the course of the training. We will distinguish the following three adaptive levels:

- **Elementary adaptive level (EAL)** – An adaptation to the static profile information of the student as name, class, type of training, the type of device to access educational resources (mobile or fixed), etc. At this level the adaptation is based on a stereotype approach. The educational resources are common to all groups of students. The adaptability of this level is realized in the phase of preparation of the typical training materials before the beginning of work in the system.

- **Static adaptive level (SAL)** – This level builds on the elementary level and is directly related to mechanisms to provide adequate learning materials for individual students according to their knowledge base, personal goals, plans and ambitions. Adaptation mechanisms are set in advance by the authors of educational resources and services, foreseeing the actions and behaviours of the typical learner. This can be realized based on the log-information about past interactions between this student and LMS and a set of rules set by the authors of the educational materials. The basic knowledge of students is determined by initial testing or by the current results from already completed training sessions. As a result, from the Lesson DB is extracted this one, which most fully meets the basic knowledge, stereotypical characteristics, objectives and plans of the individual student. Adaptability of this level is achieved before the system operation or in its initial phase when the concrete training scenario is specified. Adaptability can be improved significantly if using intelligent agents as personal assistants for each student, which will monitor and guide the entire learning process.

- **Dynamic adaptive level (DAL)** – This level complements and builds on the previous two ones. It is related to the dynamic interaction between students and the system during the training (in run-time). After selecting the most appropriate e-lesson in the previous adaptive level, LMS starts the learning process as a sequence of actions set by the author of e-content and the behaviours of the individual student. Based on the intermediate results during the training and information from previous sessions, the system adapts dynamically to the changing characteristics of the learning environment, generates new "condition-action" rules and continues the training process or starts a new more appropriate e-Lesson. At this level, in the process of dynamic interaction between learners and the training system it is essential proactive to use intelligent proactive agents, who interact with the system and with each other, so as to provide a flexible change of training scenarios, depending on the behaviours and actions of the individual student.

**SCORM** (Sharable Content Object Reference Model) [3] is a part of ADL (Advanced Distributed Learning) initiative for the creation of e-learning materials. The current version of the standard gives us priorities in following areas:

- **Reusability** – Content is independent of learning content. This reusability makes it possible to use learning materials many times, in numerous training situations, for different groups of learners with any number of development tools or delivery platforms;

- **Durability** – The content does not require modification to operate as software system and platforms are changed or upgraded;

- **Accessibility** – The created learning content could be identified and located when it is needed;

- **Interpretability** - Learning content will function in multiple applications, hardware and software configurations regardless of the tools used to create it and the platform on which it is delivered. It could be located and used according to specific training aims. One content item could be used at various difficulty levels with multiple classes and student groups;

SCORM includes three main models:

- **SCORM CAM** (Content Aggregation Model) for describing the separate components and informational items, the manner of binding to bigger structures; how the created materials are shared between two or more LMS<sup>2</sup>; the method for creation of packages and sending to LMS for execution;

- **SCORM RTE** (Run-time Environment). The model decrypts the requirements to LMS for dynamic management of the run-time environment – methods for publication, sharing and using of learning materials and resources.

- **SCORM SN** (Sequencing and Navigation) – for describing the sequencing of LMS activities during the personal training process, according to events performed and student behavior. The activities are dependent on the sequencing of rules that the LMS executes, delivering initiatives between separated learning objects.

The actual version of SCORM 2004 4th edition supplies some advantages, mainly connected with:

- Additional tools to measure progress-SCORM 2004, always includes an element of the model data to "measure progress" that shows "how complete" is individual SCO for every user. These data have already been described using different values of the measurement. This measurement and description gives an accurate picture of the overall

---

<sup>2</sup> LMS-Learning Management System



completion of the training of the user and enables the LMS to provide more precise levels of progress.

- Jump-navigation- in many courses is needed the SCOs to control navigation in a way that is different from what is available to the user. The new "Jump" navigation gives authors of educational content more options to set the sequence and navigation.

- Shared data between SCOs-the actual edition of the standard allowing the SCOs to share arbitrary amounts of data. When creating a course sequence is very convenient to have a common set of data to be accessible and usable by all SCOs

- More accessible to global data object - all object data that can be used at work are now available and can be shared with other SCOs and courses through global objects. This provides a simple and creative strategy for the implementation of educational scenarios.

### **A Model for adaptation to the student's knowledge**

Adaptation to the base knowledge of the student is associated with defining the areas of base knowledge by measuring the level of understanding of various concepts in a subject domain of the particular student. There are different approaches for the realization of this kind of adaptability. Besides the stereotypical approach, we can use the overlay model and the combined approach. The overlay model considers the user's knowledge as a subset of common knowledge, supported by the system. This model is among the dominant types of user models, usually represented as a hierarchical or semantic network of nodes directly linked to concepts from the subject domain. We can use a logical or numerical value for assessing the student's knowledge. The education in different subject domains is a prerequisite for creating of domain ontology, including basic concepts and relationships between them. To each concept there will be attributed a relative numerical value (in%), which shows the degree of certainty of the system about the student's knowledge. We use the initial test from different subjects, connected with this to determine basic knowledge of the student who is new for the system. The test results are evaluated on two levels: as a general result for the student of each subject domain and as an evaluation of the level of knowledge about each concept from ontology. The evaluation of the first level is used to determine the stereotypical group of the student and the results of the second level - to realize the model of overlapping with concepts from the domain.

The training resources and tutorials are created by authors in a special domain-based development environment. Let us concentrate our attention on two basic characteristics of the lesson - the content and structure. The content of lessons is related to specific topics, which in

turn are part of specific domains. Therefore, the e-lesson is a semantic structure of the knowledge contained in a particular area. Formally, it is an instance of a particular part of the ontology, describing the subject area, in which the individual concepts are associated with real information resources that represent them. The structure of the e-lesson depends on defined didactic goals and the characteristics of the subject area. The author of the lesson can structure the learning resources in different ways depending on their goals. As a result of studies on the structure of the lesson we came to the conclusion that we could formalize the different kinds of e-lessons according to didactic goals by creating standard scenarios for training and templates that describe them. Each template we will see as a combination of: learning resources, structure and scenario for training. To achieve this we use the basic templates and models from SCORM Best practice Guide for Content Developers and create a model for their parameterization.

The created e-Learning resources are stored in online repository. They are associated with concepts from the ontology and provide itself into the development environment for creating e-courses. The structure of lesson is determined by the author using the parameterization of some of the basic templates. Thus creates an instance of the template in which no free parameters. To conduct educational process itself must determine the training scenario. It is directly related to didactic goals, basic knowledge and behaviors of students. The authors of the lessons describe the various options and determine the rules under which will be held the learning process. Formalization of these scenarios can be realized also through the parameterization of the basic templates. Therefore, the e-Lesson will be presented in the system as a specific instance of some basic template, which (by setting values of parameters) is associated with specific learning resources. In this template are determined the structure and rules for training. Creation of educational resources (SCOs ) will not be examined in the model, as they are created as independent units, stored into special SCO-repository. They are used for creation of e-Lessons and are associated with the concepts from different ontologies.

### **Creation of the e-learning SCORM based courses**

Here will be presented creation of the SCORM e-textbook in Software engineering [7]. This e-textbook is used in Software engineering master's program in Plovdiv University in DeLC Portal [6].

Creation of the Software engineering e-textbook was done in several steps as follow:

–Designing e-textbook in SCORM format – In this step e-learning content was divided in small e-learning elements (SCOs), which have to

be independent and reusable. In this way learners will be able to adapt e-learning course, according to knowledge level of the students. They will be able to add or remove e-learning elements (SCOs) from course. To create design structure of the e-learning course in software engineering in SCORM format we use activity diagrams, which clearly present different SCORM elements and sequencing and navigation between them. These sequencing and navigation define pedagogical model in that will be presented e-learning content – learners can order which SCO will be first, which SCO will be next, when students have to pass test and so on. Figure 1 presents activity diagram of one part of the e-textbook in software engineering (Object-oriented analysis). Except design structure, figure 1 presents navigation scheme about e-learning lesson. This navigation scheme depicts management of the lesson, based on concrete lesson test. If the student passes test, in our case he has to pass correct all questions (100%), he/she continues with the next lesson in the SCORM package, otherwise depending on wrong questions, DeLC SCORM Player [7] forwards the student to the e-learning element (SCO), which consist corresponding e-learning content – marked in the navigation scheme.

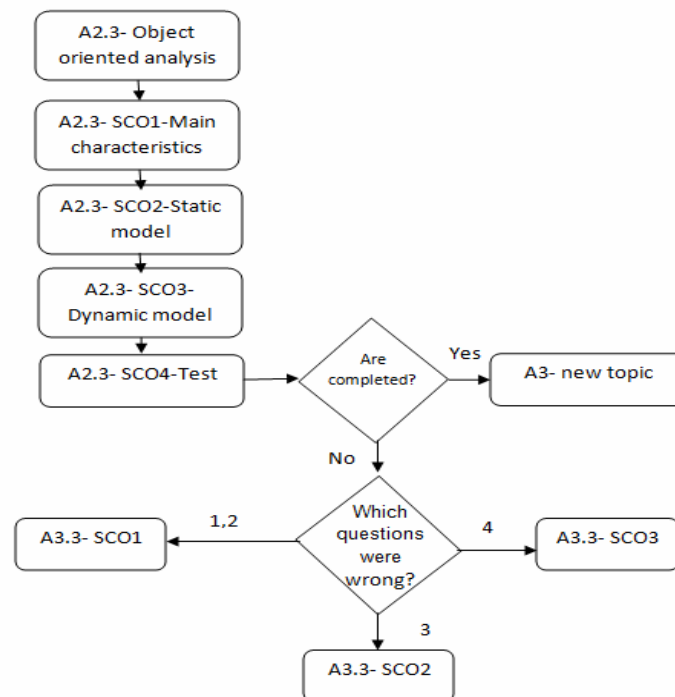


Figure1: Activity diagram of software engineering e-textbook and navigation scheme

—The second step in creation of the e-textbook in Software engineering is implementation of the designed SCORM package in SCORM Editor – for our goals we used Trident 2.0 [6] – it implemented e-learning components, order of execution of the components and it was added all e-learning resources.

—The third step of the creation of SCORM textbook in software engineering is implementation of the control tests for each theme. Control tests are implemented as different SCOs in e-learning textbook.

—And the end step is testing of the e-textbook in our DeLC SCORM Player [8] (fig. 2). Now we use the e-textbook in Software engineering master's program in Plovdiv University.

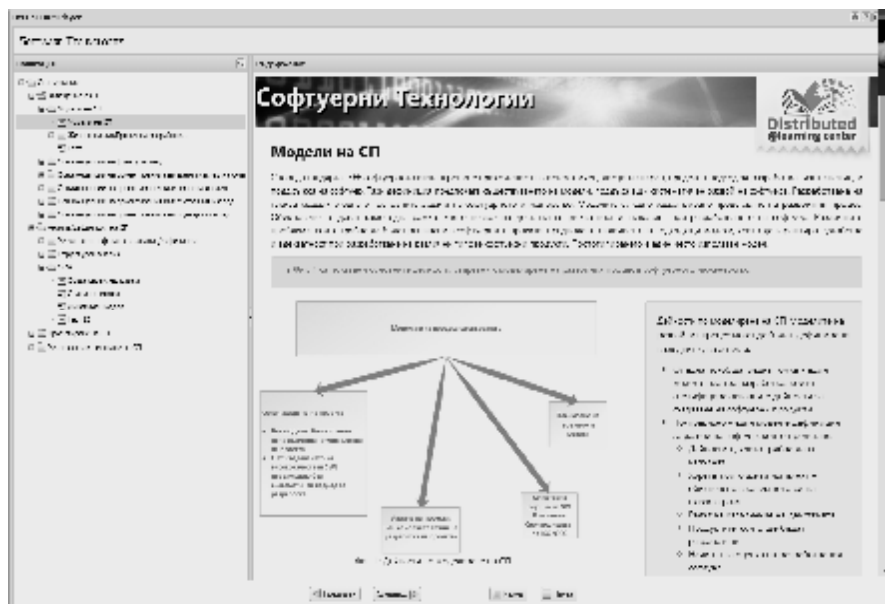


Figure 2: Software engineering SCORM textbook in DeLC SCORM Player

As an implementation tools for SCORM textbook we use - Trident IDE 2.0 [9], HTML5, CSS3, XML, JavaScript, Bootstrap and DeLC SCORM Player.

The using of the SCORM e-textbook in Software engineering increase the interests of the students in the subject and improve the educational process, because it suggest more adaptable education for the each student. Furthermore it allows collecting additional information about students:

—The lecturer can monitor the students' progress (thru DeLC)

- The lecturer can receive different statistics:
  - P** Which part of the book is most ease/difficult (based on the failed internal tests)
  - P** Which part of the book is most ease/difficult per student
- By using this model, in the portal for electronic and distance learning DeLC were created and implemented various SCORM based e-learning courses.

### **Conclusions and future work**

The presented model describes an ability to create and use SCORM-based e-Learning courses. The adaptability to user knowledge is implemented at several levels. Elementary adaptive level ensures that each student is included in the appropriate stereotype hierarchy of students by the specialty and course. This is guarantee from DeLC-portal architecture and included e-Services.

Static adaptive level is implemented at two levels: the preliminary stage and in the real learning process. After the development of a learning scenario and preparation of the SCOs, the author creates e-course in specialized SCORM- based development tool. According to learning scenario from collection of templates he chooses the right one, which after parameterization became the basis of the course structure. The author adds learning resources (SCOs) and establishes rules to be realized in the current learning scenario. After the launch of the e-course the student performs the steps of the learning scenario. SCORM LMS monitors the behavior of the student and according to rules set by the author starts next learning resource.

Dynamic adaptive level is connected with dynamic interaction between SCORM LMS and the students in the real time. When the student gave the wrong answer to the questions or need additional information, the his personal agent is associated with agents who interact with the appropriate ontology and dynamically retrieves additional information or generate questions to assist student to complete the learning course successfully.

We continue the work on the implementation an agent-oriented version of e-learning system, as well as the realization of scenarios related to adaptability in mobile learning.

### **References**

- [1] Brusilovsky, P., KnowledgeTree: a distributed architecture for adaptive elearning, In Proc. Of the 13th international World Wide Web conference on Alternate track, New York, pp. 104-113, ISBN:1-58113-912-8
- [2] Glushkova, T.; Stojanova, A.(2008) Interaction and adaptation to the specificity of the subject domains in the system for e-Learning and distance training DeLC, Proceedings of International Conference "Informatics in the

Scientific Knowledge”, pp.295-307, ISSN 1313-4345, ISBN-13:978-954-715-303-526-28, Varna, Bulgaria, June 17-19, 2008

[3] Glushkova T., M. Trendafilova, N.Uzunova, Application of SCORM standard for e-learning in the secondary school, In: Proc. Of the International Conference “Informatics in the Scientific Knowledge”, 18-22 June 2006, Varna, pp.205-216, 2006.ISSN 1313-4345, ISBN – 10:954-715-303-X, ISBN-13:978-954-715-303-5

[4] Stojanov S., I. Ganchev, I. Popchev, M. O'Droma, From CBT to e-Learning, Information Technologies and control, No 4/2005, Year III, pp. 2-10, ISSN 1312-2622.

[5] Stojanov S., I. Gancev, M. O'Droma, E. Dojchev, G. Cholakov, Agent oriented Distributed E-learning Center, International Conference, Varna, 2005, ISBN 3-929757-76-1

[6] Stoyanov S., I. Popchev, E. Doychev, D. Mitev, V. Valkanov, A. Stoyanova-Doycheva, V. Valkanova, I. Minov, DeLC Educational Portal, Cybernetics and Information Technologies, Vol. 10, No. 3., Bulgarian Academy of Sciences, 2010, pp. 49-69, ISSN 1311-9702

[7] А. Стоянова-Дойчева, В. Вълканова, Ц. Кошутански, Е. Дойчев, В. Вълканов, И. Попчев, С. Стоянов. Създаване на електронно учебно съдържание, Международна конференция „From DeLC to VelSpace”, 26-28 март 2014, Пловдив (за печат)

[8] Е. Дойчев, „Среда за електронни образователни услуги“, дисертация, ПУ „П. Хилендарски“, 2013, Пловдив

[9] Authoring tool for SCORM content development Trident 2.0 <http://www.scormsoft.com/trident> - visited on 15.05.2014

#### **Authors:**

1. **Asya Stoyanova-Doycheva**, PhD, Department of Computer Systems, University of Plovdiv “Paisii Hilendarski”, Plovdiv, Bulgaria, 0888668617, [astoyanova@uni-plovdiv.bg](mailto:astoyanova@uni-plovdiv.bg)

2. **Todorka Atanasova Glushkova**, PhD, Department of Computer Technologies, University of Plovdiv “Paisii Hilendarski”, Plovdiv, Bulgaria, 0878212950, [todorka.glushkova@gmail.com](mailto:todorka.glushkova@gmail.com)

# MATHEMATICAL METHODS IN INFORMATICS

## TRANSLATIONALLY INVARIANT SOLUTIONS OF THE HELFRICH SHAPE EQUATION

Eddie Chakarov, Vladimir Pulov, Ivailo Mladenov

**Abstract.** By making use of the conformal coordinates and applying the Lie group method we construct group-invariant solutions of the Helfrich shape equation.

**Keywords.** Symmetries of differential equations, surface shape equation, group-invariant solutions.

### 1. Helfrich Shape Equation

In the Helfrich spontaneous curvature model [3] the equilibrium shapes of the fluid membranes, such as soap bubbles, lipid vesicles, biological membranes, etc., are obtained as solutions of the *Helfrich shape equation*

$$\Delta H + 2(H^2 + \ln H - K)(H - \ln) - \frac{2\lambda H}{k} + \frac{p}{k} = 0. \quad (1)$$

The membrane is regarded as a smooth surface  $\mathcal{S} \subset \mathbb{R}^3$  and is represented by its mean and Gaussian curvatures  $H$  and  $K$ , respectively. The physical characteristics of the membrane are accounted for by the bending rigidity  $k$ , the tensile stress  $\lambda$ , the osmotic pressure  $p$  and the so called *spontaneous mean curvature*  $\ln$ . The notation  $\Delta$  is used for the Laplace-Beltrami operator on  $\mathcal{S}$ .

Over the last four decades the Helfrich shape equation (1) has attracted much attention but nevertheless only a few explicit solutions have been obtained. More about the Helfrich model and a detailed review of the solved cases can be found in [5, 6] and the references cited therein.

By adding the so called Gauss-Codazzi-Mainardi integrability conditions [8], and passing to some local coordinates on  $\mathcal{S}$ , the equation (1) is transformed to a system of partial differential equations. Here we introduce the *conformal coordinates*  $(x, y)$  and the four unknown functions [2, 4]

$$q = q(x, y), \quad \varphi = \varphi(x, y), \quad \theta = \theta(x, y), \quad \omega = \omega(x, y)$$

specifying the *conformal metric* on  $\mathcal{S}$

$$ds^2 = 4q^2 \varphi^2 (dx^2 + dy^2)$$

and the matrix of the second fundamental form

$$b = \begin{pmatrix} \theta & \omega \\ \omega & 8q^2\varphi(1 + \ln\varphi) - \theta \end{pmatrix}.$$

By making use of the formulas of the classical differential geometry [8] the mean and the Gaussian curvatures take the form

$$H = \frac{1}{\varphi} + \ln$$

$$K = \frac{1}{4q^4\varphi^4} [\varphi^2(q_x^2 + q_y^2) + q^2(\varphi_x^2 + \varphi_y^2) - q\varphi^2(q_{xx} + q_{yy}) - q^2\varphi(\varphi_{xx} + \varphi_{yy})]$$

where  $q_x \equiv \frac{\partial q}{\partial x}$ , etc. Acting with the respective Laplace-Beltrami operator

$$\Delta = \frac{1}{4q^2\varphi^2} \left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)$$

the Helfrich equation (1) and the Gauss-Codazzi-Mainardi equations are transformed to the following system of nonlinear partial differential equations [2, 9]

$$\begin{aligned} q^2(\varphi_{xx} + \varphi_{yy}) + 2q\varphi(q_{xx} + q_{yy}) - 2\varphi(q_x^2 + q_y^2) + q^4(8\varphi + \alpha_2\varphi^2 + \alpha_3\varphi^3 + \alpha_4\varphi^4) &= 0 \\ \theta_y - \omega_x - \left(8 + \frac{\alpha_2}{3}\varphi\right)q(\varphi q_y + q\varphi_y) &= 0 \\ \omega_y + \theta_x - \frac{\alpha_2}{3}q\varphi(\varphi q_x + q\varphi_x) - 8q\varphi q_x &= 0 \\ 4q^2\varphi(\varphi_{xx} + \varphi_{yy}) + 4q\varphi^2(q_{xx} + q_{yy}) - 4\varphi^2(q_x^2 + q_y^2) - 4q^2(\varphi_x^2 + \varphi_y^2) \\ - \omega^2 - \theta^2 + \left(8 + \frac{\alpha_2}{3}\varphi\right)q^2\varphi\theta &= 0 \end{aligned} \quad (2)$$

where  $\alpha_2 = 24\ln$ ,  $\alpha_3 = 8(2\ln^2 - \lambda/k)$  and  $\alpha_4 = 4p/k - 8\lambda\ln/k$ . Hereinafter, the system (2) is called the *Helfrich system*.

## 2. Lie Symmetry Group

A *Lie symmetry group* of a given system of differential equations is a Lie group of transformations of the independent and the dependent variables that leave the system invariant [7]. In the case of the Helfrich system (2) the one-parameter Lie symmetry group is of the form

$$\tilde{x}^i = \phi^i(\vec{x}, \vec{u}, \varepsilon), \quad \phi^i \Big|_{\varepsilon=0} = x^i, \quad i = 1, 2$$



$$\bar{u}^\alpha = \psi^\alpha(\vec{x}, \vec{u}, \varepsilon), \quad \psi^\alpha \Big|_{\varepsilon=0} = u^\alpha, \quad \alpha = 1, 2, 3, 4$$

where  $\varepsilon$  ( $\varepsilon \in \delta \subset \mathbb{R}$ ,  $0 \in \delta$ ) is the group parameter and the vectors  $\vec{x} = (x^1, x^2)$  and  $\vec{u} = (u^1, u^2, u^3, u^4)$  denote respectively the independent variables  $x^1 = x$ ,  $x^2 = y$  and the dependent variables  $u^1 = q$ ,  $u^2 = \varphi$ ,  $u^3 = \theta$ ,  $u^4 = \omega$ .

The method for finding the Lie symmetry group is based on the fundamental correspondence between the Lie symmetry group and the *Lie symmetry algebra* of the group generators

$$X = \sum_{i=1}^2 \xi^i(\vec{x}, \vec{u}) \frac{\partial}{\partial x^i} + \sum_{\alpha=1}^4 \eta^\alpha(\vec{x}, \vec{u}) \frac{\partial}{\partial u^\alpha}$$

where the coefficient functions  $\xi^i(\vec{x}, \vec{u})$ ,  $\eta^\alpha(\vec{x}, \vec{u})$  are solutions of the so called *determining system* [7]. The relation between the Lie symmetry group and its Lie algebra is determined by the equations (the *Lie equations*)

$$\begin{aligned} \frac{d\phi^i}{d\varepsilon} &= \zeta^i(\vec{\phi}, \vec{\psi}), & \phi^i \Big|_{\varepsilon=0} &= x^i, & i &= 1, 2 \\ \frac{d\psi^a}{d\varepsilon} &= \eta^a(\vec{\phi}, \vec{\psi}), & \psi^a \Big|_{\varepsilon=0} &= u^a, & a &= 1, 2, 3, 4 \end{aligned}$$

where  $\vec{\phi}$  and  $\vec{\psi}$  are vectors with coordinates  $\phi^i$  and  $\psi^a$ .

The starting point of the Lie group analysis of any complicated system of differential equations is the determining system. The determining system is an over-determined system consisting of linear partial differential equations. For the considered Helfrich system (2), there are 206 first and second order partial differential equations that constitute the determining system. For creating and solving of this great number of equations we took advantage of the Mathematica<sup>®</sup> package *LieSymm-PDE* [10]. We managed to do most of the symbolic calculations automatically and to reduce the determining system to 10 equations for five unknown functions of the form

$$\begin{aligned} \xi^1 &= h(x, y), & \eta^1 &= q\sigma(x, y), & \eta^3 &= v(x, y, q, \varphi, \theta, \omega) \\ \xi^2 &= r(x, y), & \eta^2 &= -C\varphi, & \eta^4 &= w(x, y, q, \varphi, \theta, \omega) \end{aligned}$$

where  $C = \text{const.}$  In order to solve the remaining part of the equations we applied the *LieSymmPDE* package in interactive mode feeding back

the program with the solutions we had found by hand. Finally we arrived at the Lie symmetry algebra of the Helfrich system for two distinguished cases (cf [2, 9]):

**Case 1.**  $|\alpha_2| + |\alpha_3| + |\alpha_4| \neq 0$

$$X^I(\xi^1, \xi^2) = \xi^1 \partial_x + \xi^2 \partial_y - q \xi_x^1 \partial_q - 2(\theta \xi_x^1 + \omega \xi_x^2) \partial_\theta - 2 \left[ \omega \xi_x^1 - \left( \theta - 4q^2 \varphi - \frac{\alpha_2 q^2 \varphi^2}{6} \right) \xi_x^2 \right] \partial_\omega$$

**Case 2.**  $\alpha_2 = \alpha_3 = \alpha_4 = 0$

$$X^{\text{II}}(\xi^1, \xi^2) = X_1(\xi^1, \xi^2) + cX_2, \quad c \in \mathbb{R}$$

$$X_1(\xi^1, \xi^2) = \xi^1 \partial_x + \xi^2 \partial_y - q \xi_x^1 \partial_q - 2(\theta \xi_x^1 + \omega \xi_x^2) \partial_\theta - 2[\omega \xi_x^1 - (\theta - 4q^2 \varphi) \xi_x^2] \partial_\omega$$

$$X_2 = \varphi \partial_\varphi + \theta \partial_\theta + \omega \partial_\omega$$

where  $\xi^1 = h(x, y)$  and  $\xi^2 = r(x, y)$  are arbitrary real-valued harmonic functions satisfying the Cauchy-Riemann conditions

$$\frac{\partial \xi^1}{\partial y} = -\frac{\partial \xi^2}{\partial x}, \quad \frac{\partial \xi^1}{\partial x} = \frac{\partial \xi^2}{\partial y}$$

and  $\partial_x \equiv \frac{\partial}{\partial x}$ , etc. The full sets of group generators  $X^I(\xi^1, \xi^2)$  and

$X^{\text{II}}(\xi^1, \xi^2)$  constitute two infinite dimensional symmetry Lie algebras.

### 3. Translationally Invariant Solutions

A *group-invariant solution* of a given system of differential equations is a solution that remains unaltered under the action of the Lie symmetry group. The group-invariant solutions are obtained by solving the so called *reduced system* – a system of differential equations that has a fewer number of independent variables compared with the originally considered system [7].

We look for solutions of the Helfrich system (2) that are invariant under the group of the translations of the variable  $x$ , i.e., solutions invariant under the Lie symmetry group generated by the subalgebra of the general Lie symmetry algebra spanned by the Lie group generator  $X^I(1, 0) \equiv \partial_x$ . The symmetry reduction technique shows that the translationally invariant solutions have the form

$$q = q(y), \quad \varphi = \varphi(y), \quad \theta = \theta(y), \quad \omega(y) = \alpha_5 = \text{const}$$

and can be obtained from the reduced system of ordinary differential equations (cf [2])

$$\begin{aligned}
q^2 \varphi_{yy} + 2q\varphi q_{yy} - 2\varphi q_y^2 + q^4 (8\varphi + \alpha_2 \varphi^2 + \alpha_3 \varphi^3 + \alpha_4 \varphi^4) &= 0 \\
\theta_y - q \left( 8 + \frac{\alpha_2}{3} \varphi \right) (\varphi q_y + q \varphi_y) &= 0
\end{aligned} \quad (3)$$

Searching for solutions of (3) we restrict to the case when the two functions  $q(y)$  and  $\varphi(y)$  satisfy the relation

$$q(y) = \frac{1}{\varphi(y)}.$$

Under this assumption we readily obtain

$$\theta(y) = 0, \quad \omega(y) = 0$$

$$\int \left( C_1 \varphi^4 - 2\alpha_4 \varphi^3 - \alpha_3 \varphi^2 - \frac{2}{3} \alpha_2 \varphi - 4 \right)^{-\frac{1}{2}} d\varphi = y + C_2$$

where  $C_1$  and  $C_2$  are arbitrary real constants. The latter integral is expressible in terms of elliptic functions or in some special cases in elementary functions. Let us first assume that the quartic polynomial

$$f(\tau) = C_1 \tau^4 - 2\alpha_4 \tau^3 - \alpha_3 \tau^2 - \frac{2}{3} \alpha_2 \tau - 4$$

has no repeated factors and the coefficients  $C_1$  and  $\alpha_4$  are not equal to zero. Then we can make use of the Weierstrassian form of the integral (see e.g. [12])

$$\int_{\zeta}^{\infty} \frac{dt}{\sqrt{4t^3 - g_2 t - g_3}} = y + C_2 \quad (4)$$

where

$$g_2 = \frac{1}{12} (\alpha_2^2 - 4\alpha_2 \alpha_4 - 48C_1), \quad g_3 = \frac{1}{216} (\alpha_2^3 - 6\alpha_2 \alpha_3 \alpha_4 + 216\alpha_4^2 - 6C_1 \alpha_2^2 + 144C_1 \alpha_2)$$

are the so called invariants of the polynomial  $f(\tau)$ . The lower limit of the integral is

$$\zeta = \frac{f'(a)}{4(\varphi - a)} + \frac{f''(a)}{24} \quad (5)$$

where  $a$  is an arbitrary root of the polynomial and

$$f'(a) \equiv \frac{df(\tau)}{d\tau} \Big|_{\tau=a}, \text{ etc. The equation (4) is equivalent to the relation [1, 12]}$$

$$\zeta = \wp(y + C_2; g_2, g_3) \quad (6)$$

where  $\wp(y + C_2; g_2, g_3)$  is the Weierstrassian elliptic  $\wp$ -function. Combining the equations (5) and (6) we obtain the solution

$$\varphi(y) = \alpha + \frac{f'(\alpha)}{4} \left( \wp(y + C_2; g_2, g_3) - \frac{f''(\alpha)}{24} \right)^{-1}.$$

In the cases when the coefficients of the polynomial  $C_1$  and  $\alpha_4$  vanish,  $C_1 = \alpha_4 = 0$ , or the polynomial has repeated factors (if the equation  $f(\tau) = 0$  has multiple roots) the solution can be expressed in terms of elementary functions. Several examples of the latter case are presented in the Table 1.

Table 1. Solutions in terms of elementary functions

$C_1$	$\alpha_2$	$\alpha_3$	$\alpha_4$	$f(\tau)$	$\varphi(y)$
$-4$	$\pm 24$	$24$	$\pm 8$	$-4(\tau \pm 1)^4$	$\frac{i}{2(y + C_2)} m1$
$-1$	$\pm 12\sqrt{2}$	$12$	$\pm 2\sqrt{2}$	$-(\tau \pm \sqrt{2})^4$	$\frac{i}{y + C_2} m\sqrt{2}$
$-1$	$\pm \frac{39}{2}$	$15$	$\pm \frac{7}{2}$	$-(\tau \pm 1)^2(\tau \pm 4)$	$\frac{(3y + 3C_2 \pm 2i)(y + C_2 - m2i)}{m(y + C_2)(3y + 3C_2 - m4i)}$
$-1$	$\pm 18$	$13$	$\pm 3$	$-(\tau \pm 1)^2(\tau \pm 2)^2$	$\mp \frac{3}{2} + \frac{i}{2} \cotg\left(\frac{y + C_2}{2}\right)$
$4$	$12$	$0$	$-4$	$4(t-1)(t+1)^3$	$-\frac{2(y + C_2)(y + C_2 - 1) + 1}{2(y + C_2)(y + C_2 - 1)}$
$4$	$-12$	$0$	$4$	$4(\tau - 1)^3(\tau + 1)$	$\frac{2(y + C_2)(y + C_2 + 1) + 1}{2(y + C_2)(y + C_2 + 1)}$

The next two solutions depend on two integration constants  $C_1 > 0$  and  $C_2 \in \mathbb{R}$  and generalize (via  $C_1$ ) the real-valued solutions given in the last two rows of the Table 1.

First we assume  $\alpha_2 = \frac{3(12 - C_1)}{2}$ ,  $\alpha_3 = 3(4 - C_1)$ ,  $\alpha_4 = \frac{4 - 3C_1}{2}$ , getting the solution

$$\varphi(y) = \frac{16C_1 + (\sqrt{C_1}(C_1 + 4)(y + C_2) - 2C_1)^2}{4C_1^2 - (\sqrt{C_1}(C_1 + 4)(y + C_2) - 2C_1)^2}.$$

This solution is related to the polynomial  $f(\tau) = (C_1\tau - 4)(\tau + 1)^3$  and for  $C_1 = 4$  coincides with the solution in the penultimate row of the Table 1.

The second solution

$$\varphi(y) = \frac{16C_1 + (\sqrt{C_1}(C_1 + 4)(y + C_2) + 2C_1)^2}{(\sqrt{C_1}(C_1 + 4)(y + C_2) + 2C_1)^2 - 4C_1^2}$$

is obtained on writing  $\alpha_2 = \frac{3(C_1 - 12)}{2}$ ,  $\alpha_3 = 3(4 - C_1)$ ,  $\alpha_4 = \frac{3C_1 - 4}{2}$ , and for  $C_1 = 4$  coincides with the solution in the last row of the Table 1. The polynomial that corresponds to this solution is  $f(\tau) = (\tau - 1)^3(C_1\tau + 4)$ .

#### 4. Conclusion

In this paper the Helfrich spontaneous curvature model has been considered from group-theoretical viewpoint. The Lie group method has been applied to the conformal metric representation of the Helfrich model. By applying the Mathematica® *LieSymm-PDE* package the Lie symmetry algebra of the Helfrich system has been obtained. A symmetry reduction technique has been used to construct translationally invariant solutions to the Helfrich system. The equilibrium shapes of the fluid membranes corresponding to the real-valued solutions, expressed by the Weierstrassian elliptic  $\wp$ -function, have been discussed elsewhere [11]. Let us note that the complex-valued solutions, though mathematically correct, are not directly related to the considered physical model and the geometry of the membranes.

#### References

- [1] Abramowitz M., Stegun I. (Eds.), Handbook of Mathematical Functions, New York, Dover, 1972.
- [2] De Matteis G., Manno G., Lie Algebra Symmetry Analysis of the Helfrich and Willmore Surface Shape Equations, *Comm. Pure Appl. Anal.* **13** (2014), 453-481.

- [3] Helfrich W., Elastic Properties of Lipid Bilayers: Theory and Possible Experiments, *Z. Naturforsch C* **28** (1973), 693-703.
- [4] Konopelchenko B., On Solutions of the Shape Equation for Membranes and Strings, *Physics Letters B* **414** (1997), 58-64.
- [5] Mladenov I., New Solutions of the Shape Equation, *Eur. Phys. J. B* **29** (2002), 327-330.
- [6] Mladenov I., Hadzhilazova M., Djondjorov P., Vassilev V., Some Explicit Solutions of the Shape Equation, *AIP Conference Proceedings* **1301** (2010), 215-227.
- [7] Olver P., Applications of Lie Groups to Differential Equations, Berlin, Springer, 1986.
- [8] Oprea J., Differential Geometry and its Applications, Washington D. C., Mathematical Association of America, 2007.
- [9] Pulov V., Chakarov E., Hadzhilazova M., Mladenov I., Symmetry Properties of the Membrane Shape Equation, *Geometry, Integrability & Quantization* **14** (2013), 152-159.
- [10] Pulov V., Chakarov E., Uzunov I., A Computer Algebra Application to Determination of Lie Symmetries of Partial Differential Equations, *Serdica J. Computing* **1** (2007), 241-251.
- [11] Pulov V., Hadzhilazova M., Mladenov I., Cylindrical Helfrich Surfaces, In: Proceedings of the Forty Third Spring Conference of the Union of Bulgarian Mathematicians, Borovetz, April 2-6, 2014, Sofia 2014, pp. 132-137.
- [12] Whittaker E., Watson G., A Course of Modern Analysis, Fourth Edition, Cambridge, University Press, 1996.

#### Authors:

**Eddie Chakarov**, Assoc. Prof., PhD, Department of Computer Science, Varna Free University, Chaika Resort, 9007 Varna, Bulgaria  
Phone: +359 52 781 001, e-mail: [eddy@abv.bg](mailto:eddy@abv.bg)

**Vladimir Pulov**, Assist. Prof., PhD, Department of Physics, Technical University of Varna, Studentska Str. 1, 9010 Varna, Bulgaria  
Phone: +359 52 383 285, Email: [vpulov@hotmail.com](mailto:vpulov@hotmail.com)

**Ivailo Mladenov**, Assoc. Prof., PhD, Department of Lipid-Protein Interactions in Biological Membranes, Institute of Biophysics and Biomedical Engineering, Bulgarian Academy of Sciences, Acad. G. Bonchev Str., Bl. 21, 1113 Sofia, Bulgaria, Phone: +359 2 9 792 637, e-mail: [mladenov@obzor.bio21.bas.bg](mailto:mladenov@obzor.bio21.bas.bg)

## DISCOVERING ISSUES IN CODE USING CALL AND REFERENCE GRAPHS

Todor Cholakov, Dimitar Birov

**Abstract:** *During the evolution of a software product its code changes and becomes more and more complicated. As a result there are chances for bugs and problems to appear. In order to discover potential issues in the program code there have been developed heuristic algorithms also known as "bad smell". In the current paper we discuss approaches for using call and reference graphs in order to discover some of the most often issues in program code - duplicate code, god classes and classes suitable for splitting into several smaller classes.*

**Keywords:** *refactoring, reengineering, code patterns, call graphs, reference graphs*

### Introduction

In the process of developing a software product, its code keeps changing all the time. In most cases the changes are introduced by different developers having different levels of knowledge about the software structure. As a result it often happens that there are functions in different parts of the product that are nearly exact copies of each other, some classes become more and more complicated over time and some implement the logic for more than one functionality. Code like this becomes prone to errors and is hard to read and maintain [2].

Issues like the ones we stated above are discovered by the programmers either by accident or when they are performing a refactoring of the product. However in most cases such issues remain unnoticed or even if someone discovers them no steps are performed to resolve them because "This module is too complicated and problematic and I don't know what will get broken if I change this part".

The purpose of the current paper is to introduce algorithms for efficient discovery of some of the most widespread issues in code and in the same time to give suggestions for concrete changes in the code that will lead to its improvement. In order to achieve our task two approaches for representing program code – as a call graph and as a reference graph. In the next paragraphs we are going to give some basic definitions related to the terminology used.

### Call graph

A call graph for a software product or some part of it is a directed graph, having as nodes the methods or functions in the code and as arcs the calls between them. This kind of graphs may vary depending on the way we define the arcs:

- Canonical call graph is a call graph where there is a corresponding graph for each call, labeled with the place where the call happens. In this case it is possible to have multiple arcs from one node to another and we are able to retrieve the sequence of calls out of the graph.

- Statistical call graph is a call graph where each arc is labeled with the number of the calls performed. This kind of call graph is more suitable for performing some kind of statistical analysis.

The call graphs may vary depending on the nodes included:

- The simplest approach is to add as nodes only the explicitly defined method. However this will cause some information loss.

- Additionally the implicit and explicit constructors may be added as nodes of the graph. In this case we must take into account that the initialization of all the local fields in the class is considered part of the constructor.

- The most complicated approach is to add all the fields of the class as special method pairs – one for getting the value and one for assignments. This adds an important level of detail to the graph allowing the analyzing program to achieve a better understanding of the code.

#### **Reference graph**

A reference graph for a software product is a directed graph whose nodes are classes and the arcs are the interconnections between them. There are much more constructions which add arcs to the graph, compared to the call graphs:

- In the body of class A there is a call to method belonging to class B

- In the body of class A there is a field or variable of type B

- Class A extends class B (this includes the ***implements*** construction in Java)

- Some of the methods called in the body of A return a result of type B

- Some of the methods defined or declared in class A return result of type B

- In the body of class A there is a reference to a field of type B

- In the body of class A there is a cast to B.

- In the body of A there is a creation of new object of type B.

Each arc of a reference graph is labeled with the number of the usages between the edges.

#### **Graph clustering**

Clustering is the process of dividing the graph into subgraphs (clusters), each of them containing part of the nodes of the original graph. The nodes within a cluster must have more connections to nodes in the same cluster than to nodes belonging to other clusters. In most cases



there are multiple ways to cluster a graph and it is often difficult to say which of them is optimal or even which of them is better.

### **Discovering duplicate code**

Usually there is a lot of duplicate code in each large project. There are several causes that lead to creation of duplicate code:

- Similar functionalities are needed in different parts of the product. It is a common case that each developer in the team has knowledge only for small parts of the product code and doesn't know that the needed functionality is already implemented.
- A method or part of it is copied in another class or module without considering some other ways to reuse the code.
- Two pieces of code were designed as being different, but over time they become similar.

The complexity of the task that we are solving depends on the definition of the term "duplicate code" which we chose. There several definitions that may be considered:

1. We define two pieces of code as duplicate if after we remove all whitespace and comments where possible and replace them with a single space on all other places, the resulting code is the same. This is the simplest and yet the most restrictive definition, because there are many cases in which the code has some insignificant differences and in this case it won't be considered duplicate.

2. We define two pieces of code as duplicate, if after the transformation described above it is possible by replacing some or all of the identifiers in the first piece of code, to make it equal to the second piece of code. This definition covers the cases when the two pieces of code are equivalent to within a variable name, but also covers the cases where different methods are called.

3. We define two pieces of code as duplicate, if large consecutive parts of them are duplicate code in the sense of the previous definition.

We must be aware that duplicate code doesn't necessarily mean equivalent code. For example if two methods have the same logic, but one of them use "`System.out.println()`" for printing and the other uses "`output.println()`". Obviously they don't do exactly the same thing. However they might be considered duplicate code and after some transformations we may extract the differences in separate methods.

Our purpose in this paper is to propose an algorithm for discovering duplicate methods using the second of the above definitions.

The algorithm that we propose is the following:

1. Create a statistical call graph for the whole product. The arcs are denoted only by the explicitly called methods. The nodes of the graph are numbered.

2. For each of the nodes of the call graph we define an identifier consisting of the frequencies of the calls to its neighbors in descending order, followed by the numbers of the corresponding nodes. The list of all identifiers is sorted in ascending order.

3. Each pair of identifiers that are next to each other in the list or have the same list of frequencies, are potentially duplicate code. If two identifiers have the same lists of frequencies, but different nodes, they might be duplicate code in the sense of the second of the definitions above. If they have both equal lists of frequencies and equal lists of nodes, they may represent duplicate methods in the sense of the first definition.

4. We create a canonical call graph for the analyzed code, including as nodes both the constructors and the fields.

5. The potentially duplicate pieces of code, discovered in step 3, are analyzed using the canonical call graph in the following way:

a) We sort the outgoing arcs of the two nodes in the order of their occurrence in the code.

b) We cycle through the lists and for each pair of target nodes do the following:

i. If the frequencies of the two nodes are the same and none of them has an additional number assigned, we assign the same number to both of them.

ii. If the frequencies are the same, but one of the target nodes has an additional number assigned and the other node has no additional number or it is different from the first one, we state that the two pieces of code are not duplicate code.

iii. If the frequencies are not the same, the two pieces of code are not the same.

c) If at point b) we didn't find reason to state that the nodes don't represent duplicate code, we remember the pair of nodes and continue to the next.

6. The pairs of nodes, which we got in the previous step, are proposed for refactoring.

This algorithm is able to discover the following cases of duplicate code:

1. All cases in the sense of definition 1 are found as duplicate code  
2. All cases according to definition 2 where the same identifier must be replaced only once with another one in order, are also discovered.

3. The only cases according to definition 3, which are discovered, are the ones where the code between the pieces doesn't perform a method call or reference to a field.

A disadvantage of the proposed algorithm is that some methods are marked as duplicate code, although they are not. For example the arithmetic and logical operators, as well as conditional and cyclic operators are not taken into account by the algorithm. If we want to improve the accuracy, we may recognize these operators as special functions both in step 1 and step 4. This will reduce the number of false positives, but will also reduce the number of duplicates according to definition 3, that are recognized.

### Discovering god-classes

A god-class is a class, which has too many functions and tries to do a lot of different things. As a rule such classes concentrate too many methods and data in themselves.

We propose the following algorithm for discovering this kind of classes:

1. We create a statistical call graph for the analyzed product and we include both the constructors and the used fields as nodes in the graph.
2. We cluster the graph, using a suitable algorithm. There is a variety of clustering algorithms described in [1],[3] and [5]. Alternatively in this step we may use a reference graph, which gives better accuracy for the clustering. In both cases we need class level results.
3. We change the weights of all arcs to 1.
4. We unite all the nodes belonging to the same class in the following way:
  - a. Each node is labeled with the number of nodes which it contains.
  - b. The arcs ending in one of the contained nodes are redirected to end at the container node.
  - c. The arcs starting from some of the contained nodes are redirected to start from the container node.
  - d. The arcs starting and ending in the same container node are replaced by a single arc, having their number assigned as weight.
  - e. All arcs having the same pairs of start and end nodes are replaced by a single arc, having their count as a weight.
5. We walk through the nodes of the graph and remember these nodes, which satisfy the following conditions in a separate list:

$$wt(A_{inner}) > \sum_{x \in A_{out}} wt(x) \quad (1)$$

$$\frac{wt(A_{inner})}{|A|} > 2 \quad (2)$$

$$\frac{|A_{out}|}{|clusterOf(A)|} \geq outgoingLimit \quad (3)$$

Where:

$wt(x)$  is the weight of arc x.

$A_{inner}$  is the cyclic arc for the node A.

$A_{out}$  is the set of outgoing arcs of node A.

$|A|$  is the number of elements contained in A.

The function  $clusterOf(x)$  returns the set of nodes belonging to the same cluster as X.

$outgoingLimit$  is a parameter to the algorithm. The default value is 0,7

6. We propose for refactoring those nodes in the previous list, that have more than  $elementLimit$  elements (methods and fields), where  $elementLimit$  is a parameter to the algorithm, having default value of 40.

The conditions (1) and (2) ensure that the discovered classes are not simple dispatchers, but really do some work. The dispatchers are generally allowed to have a lot of outbound connections and a lot of methods.

The condition (3) ensures that the discovered classes dominate in their module. However this condition is not strict. If we set  $outgoingLimit$  to zero the condition will be ignored. In this case the algorithm will discover all complicated classes,

### Discovering classes suitable for splitting

The main characteristic of the classes referred in this section is that they implement the functionality for several different and independent jobs at the same time. They can easily be split into several smaller classes. The difference between them and god-classes is that the god-classes usually have large degree of connectivity to all the classes in their module, while in this case this is not compulsory.

The algorithm that we propose for discovering such classes in the analyzed product is the following:

1. We walk through all the classes in the product and perform the following steps for each of them:

a. We create a statistical call graph for the current class, which includes field references but does not include any arcs that start or finish outside the current class.

b. The resulting graph is clustered using a suitable algorithm.

c. If some of the clusters contain up to three elements, we exclude them from the set of clusters and if they don't have any connections outside connections add them to a list of potentially dead code.

d. If there is only one cluster left after the previous step, the class is not suitable for splitting and the algorithm continues to the next class.

e. If there are two or more clusters, it is added to the list of classes suitable for splitting in smaller classes

As a result of the execution of this algorithm we get a list of classes suitable for splitting in smaller parts and a suitable partitioning for each of them. As a side effect we also get a list of elements that are potentially either dead code or code that is in the wrong place. For example these might be fields and their getter and setter methods, which are never used in the rest of the class.

### **Improvements and next steps**

All of the presented algorithms may be improved. An important fact to be noted is that all of them end to the place where a proposal for suitable changes in the code is made. Important next step is to improve the last step of the analysis in order to decrease the number of wrong proposals to the possible minimum. This will allow the proposed refactorings to be carried out automatically on the whole product.

### **Conclusion**

Call and reference graphs have a lot of applications for analyzing source code. The proposed algorithms demonstrate their usage in the process of discovering potential issues in the source code. An important feature of all the algorithms is that they have a potential for being highly efficient, because usage of call and reference graphs hides most of the details and makes the analysis much simpler.

### **Acknowledgments**

The authors gratefully acknowledges financial support by the Bulgarian National Science Fund within project DO 02-102/23.04.2009.

### **References**

- [1] Cholakov, T., Birov D. A compound algorithm for clustering software, Mathematics and Education in Mathematics, Proceedings of the Forty third Spring Conference of the Union of Bulgarian Mathematicians, April 5-9, 2011, Borovetz, pp. 155-167, ISSN: 1313-3330.
- [2] Fowler, Martin (1999). Refactoring. Improving the Design of Existing Code. Addison-Wesley. ISBN 0-201-48567-2.
- [3] Praditwong, K.; Harman, M.; Xin Yao. "Software Module Clustering as a Multi-Objective Search Problem," Software Engineering, IEEE Transactions on, vol.37, no.2, pp.264,282, March-April 2011
- [4] S. Mancoridis, B. Mitchell, Y. Chen, and E. Gansner. "Bunch: a clustering tool for the recovery and maintenance of software system structures,"

in Proceedings of the International Conference on Software Maintenance (ICSM '99), IEEE Computer Society Press, Oxford, UK, August 1999.

[5] V. Tzerpos, R. C. Holt. ACDC: an algorithm for comprehension driven clustering, in: Proceedings of the Seventh Working Conference on Reverse Engineering (WCRE'00), 2000, pp. 258–267.

**Authors:**

**Todor Plamenov Cholakov**

Faculty of Mathematics and Informatics, University of Sofia  
5, James Bourchier Blvd., 1164 Sofia, Bulgaria  
e-mail: [todortk@abv.bg](mailto:todortk@abv.bg)

**Dimitar Yordanov Birov**

Faculty of Mathematics and Informatics, University of Sofia  
5, James Bourchier Blvd., 1164 Sofia, Bulgaria  
e-mail: [birov@fmi.uni-sofia.bg](mailto:birov@fmi.uni-sofia.bg)

## TWO-PHASE REGRESSION WITH SMOOTH TRANSITION

Stoyan Nenkov

**Abstract:** *A method of estimating parameters of two-phase nonlinear regression with smooth transition between phases is described. It consists of two stages, both utilizing the least square fit. In the first one, each phase is fitted separately and, simultaneously, a transition point is determined. In the second stage, the two phases are joined smoothly by a proper transition function, which depends on the transition point chosen by the grid search. The practical aspects of the method proposed are demonstrated on the data concerning the soil bulk density in dependence on the soil water content.*

**Key words:** *linear regression; nonlinear regression; two-phase regression; transition point; transition function; least square method.*

### 1. INTRODUCTION

The method of multi-phase regression needs to be used when the dependent variable behaves differently in subsequent intervals of the independent variable.

Numerous such cases have been described in literature (Sprent, 1961; Hudson, 1966; Hinkley, 1969; Bacon and Watts, 1971; Kirby, 1974; Lerman, 1980; Serber and Wild, 1989). Sometimes the need of multi-phase regression becomes apparent only after the observations are plotted in the coordinate system. There are also issues, in which it is known, prior to the experiment, that in the observed material a structural change occurs, when the independent variable exceeds a given threshold. The situation is relatively simple when the threshold value, also referred to as the transition point, is known. However, it is usually not the case. Moreover, the experimenter often expects that the transition between phases will be continuous and smooth. In the literature there are known methods of determining a multi-phase linear regression with smooth transition (Seber and Wild, 1989; Lerman, 1980) a multi-phase nonlinear regression with continuous transition, but without smoothness (Lerman, 1980), as well as the methods of smoothing the transitions between phases of linear regression (Serber and Wild, 1989). Serber and Wild also give the methods of smoothing multi-phase nonlinear functions. But they use the reparametrization which changes the interpretation of parameters involved.

The aim of this paper is to present a procedure leading to continuous and smooth two-phase linear or nonlinear regression with an unknown transition point. This procedure uses methods known earlier which, however, were significantly modified to eliminate nonuniqueness of solutions and to ensure smoothness of the resulting regression function.

Practical aspects of the approach proposed are discussed with reference to the data concerning the soil bulk density in dependence on the soil water content.

## 2. FORMALIZATION OF THE PROBLEM

The function of a two-phase regression may be presented as

$$f(x, b) = \begin{cases} f_1(x, b_1), & x \leq a, \\ f_2(x, b_2), & x > a, \end{cases} \quad (1)$$

Where  $a$  is an unknown transition point, while  $f_1$  and  $f_2$  are specific functions dependent on the unknown parameter vectors  $b_1$  and  $b_2$ . To ensure the continuity of the function (1), phases  $f_1$  and  $f_2$  must fulfill the condition

$$f_1(a, b_1) = f_2(a, b_2) \quad (2)$$

Modeling a given phenomenon with two-phase regression consists in estimating unknown parameters, i.e. the transition point  $a$  and vectors the  $b_1$  and  $b_2$ , on the basis of observations  $(x_i; y_i), i = 1, \dots, n$  where  $x_i$  is the  $i$ -th observation of the independent variable,  $y_i$  is the observation of the dependent variable corresponding to  $x_i$ , and  $n$  is the total number of observations of the investigated process. Adapting the least squares method, it is possible to seek estimates of  $a$ ,  $b_1$  and  $b_2$ , which minimize the sum

$$S = \sum_{i=1}^n (y_i - f(x_i, b))^2$$

In case of the function (1), the sum  $S$  can be expressed as

$$S = \sum_{x_i \leq a} (y_i - f_1(x_i, b_1))^2 + \sum_{x_i > a} (y_i - f_2(x_i, b_2))^2 \quad (3)$$

If the transition point is known, the minimum of  $S$  can be found minimizing each component of (3) separately, on the basis of properly divided set of observations. This, however, does not ensure the continuity of the solution. As was noted by Lerman (1980), in case of the unknown transition point and under the condition (2), it is practically impossible to solve this problem. Therefore he suggested a grid search method, which



applied to a given interval  $(a_{\min}; a_{\max})$  gives the most suitable point  $a$ . In results, the estimates of unknown parameters for both phases, can be found. However, this procedure does not guarantee the smoothness of estimated regression. The problem how to modify the obtained phases in order to achieve continuity and smoothness is discussed in the next Section.

### 3. SMOOTH TRANSITION BETWEEN PHASES

In many cases it is required that the regression function should be continuous and smooth throughout its domain. The condition (2) ensures continuity, but it does not ensure smoothness of the function at the point  $x = a$ . A sufficient condition for smoothness of (1) is the equation  $f_1'(a, b_1) = f_2'(a, b_2)$ , but it imposes too strong restrictions on the functions  $f_1$  and  $f_2$  and it may not be always used.

Another solution to the problem of continuity and smoothness of two-phase regression can be found by joining phases with the use of an additional transition function. This function should change phases to the smallest possible extent and only in the area closest to the transition point. When both phases of the function (1) are linear, Griffith and Miller (1973) and Serber and Wild (1989) suggested a transition function, denoted here by  $h$ , which fulfills the following conditions:

$$\forall_g \lim_{x \rightarrow \pm\infty} (xh(x, g) - |x|) = 0 \quad (4)$$

$$\forall_x \lim_{g \rightarrow 0} h(x, g) = \text{sgn}(x) \quad (5)$$

$$\forall_g h(0, g) = 0 \quad (6)$$

where  $g > 0$  is an unknown parameter. Then the two phases may be connected in the form

$$f(x, b) = \frac{1}{2} [f_1(x, b_1) + f_2(x, b_2) + \{f_2(x, b_2) - f_1(x, b_1)\}h(x - a, g)] \quad (7)$$

which is a continuous and smooth function. The condition (4) guarantees that the function (7) will be consistent with the phase  $f_1(x, b_1)$  to the left of the transition point and consistent with the phase  $f_2(x, b_2)$  to the right of that point. From the condition (5) it follows that the goodness of fit increases when  $g$  approaches zero. The condition (6) implies that the

function (7), at the transition point  $a$ , is the arithmetic mean of  $f_1(x, b_1)$  and  $f_2(x, b_2)$ . If the phases fulfill the condition (2), the function (7) crosses also the point of their intersection. The conditions (4)-(6) are fulfilled for example by hyperbolic tangent, i.e.  $\tanh\left(\frac{x}{g}\right)$ . Serber and Wild (1989) mentioned also a possibility of using the transition function for nonlinear regressions. However, when at least one of the phases  $f_1$  and  $f_2$  is nonlinear, the conditions (4)-(6) are not sufficient, in general.

In order to formulate conditions ensuring proper joining of two nonlinear phases, we use the convex combination of the form

$$f(x, b) = [1 - h(x - a, g)]f_1(x, b_1) + h(x - a, g)f_2(x, b_2) \quad (8)$$

This function is a smooth approximation of the two-phase regression, if the transition function  $h(x, g)$  is continuous and smooth, i.e. it possesses a continuous first derivative for all  $x \in R$ , and if it fulfills the following conditions

$$\forall_{g>0} \lim_{x \rightarrow \infty} (1 - h(x, g))f_1(x, b_1) = 0 \quad (9)$$

$$\forall_{g>0} \lim_{x \rightarrow \infty} (1 - h(x, g))f_2(x, b_2) = 0 \quad (10)$$

$$\forall_{g>0} \lim_{x \rightarrow -\infty} h(x, g)f_1(x, b_1) = 0 \quad (11)$$

$$\forall_{g>0} \lim_{x \rightarrow -\infty} h(x, g)f_2(x, b_2) = 0 \quad (12)$$

$$\forall_{x \neq 0} \lim_{g \rightarrow 0} h(x, g) = \begin{cases} 1, & x > 0 \\ 0, & x < 0 \end{cases} \quad (13)$$

$$\forall_{g>0} h(x, g) \text{ is non-decreasing in its domain} \quad (14)$$

The conditions (9)-(12) imply convergence of (8) to the phase  $f_1(x, b_1)$  on the left of the point  $a$ , and to the phase  $f_2(x, b_2)$  on the right of this point. The condition (13) ensures the convergence of function (8) to the function (1) throughout the whole domain except the transition point, and condition (14) guarantees gentle and rounded joining between phases. Since phases  $f_1$  and  $f_2$  are involved in conditions (9)-(12), it is

not possible to determine one universal transition function applicable for any two-phase regression. Nevertheless, these conditions ensure continuous and smooth joining of two phases. If the phases additionally fulfill the continuity condition (2), then the function (8), at the transition point, is equal to the common value of phases. If the condition (2) is not met, the transition point is the weighted mean of the phases at this point. It should be noted that a large difference between phases in the transition point  $a$  results in an increasing effect of one phase on the other in the neighborhood of  $a$ , which may deprive vectors  $b_1$  and  $b_2$  of their physical interpretation near the transition point.

#### 4. ESTIMATION

In accordance with the discussion presented above, the estimation of parameters of nonlinear two-phase regression will be performed in two stages. In the first one we will search for phases and the interval containing the transition point. In the second stage the two phases will be joined estimating the transition point and the parameter of the transition function.

Since the selection of the transition function depends on the phase functions, it is necessary at first to determine the forms of phases and their parametrization. The process of estimating the parameters of the first phase can be carried out sequentially. It begins by estimating the parameters of this phase on the basis of all observations, and next, in the following  $k$ -th step, on the basis of  $n - k$  first observations. This process should be continued as long as the determination coefficient  $R^2$  increases or the sum  $S$  decreases. However, it should be remembered that  $R^2$  does not always increase monotonically, especially in the initial steps. The step, in which the biggest value of the determination coefficient is obtained, gives estimates of parameters for the first phase. Simultaneously the interval  $[x_1; x_{n-k}]$  determines the domain of this phase.

Repeating this process for the second phase, this time with the elimination of consecutive initial observations, gives, in the  $l$ -th step, the estimates of parameters for the second phase and the approximation of its domain  $[x_l; x_n]$ . In consequence it may be expected that the transition point belongs to the interval  $[a_{\min}; a_{\max}]$ , where  $a_{\min} = \min(x_{n-k}; x_l)$  and  $a_{\max} = \max(x_{n-k}; x_l)$ .

In the second stage, the most suitable transition point in the interval  $[a_{\min}; a_{\max}]$  is determined by the grid search method proposed by

Lerman (1980). For each of the potential transition points, the estimate of  $g$  is evaluated. We choose the transition point and the estimate of the parameter  $g$ , corresponding to the highest value of  $R^2$ . During the second stage, the phases can be treated as fixed or their parameters can be re estimated together with the parameter  $g$ . In the latter case it is necessary to control the effect of the selection of the transition point, as it may have large influence on the phases.

## 5. APPLICATION

The method proposed was developed during the research on soil mechanics, when soil bulk density was analyzed in dependence on compaction soil water content. It was observed that the reaction of bulk density to changes in soil water content is not uniform. In the initial stage, bulk density increases exponentially along with the increase of water content in soil. After the maximum is reached, at the point called the optimum compaction water content or standard water content, it decreases approaching values determined by the line of the maximum pore filling with water. In result, the bulk density was modeled with the use of a two-phase regression function. Its first phase is described by an exponential function of the form

$$f_1(x, b_1) = b_{11} + e^{b_{12} + xb_{13}}$$

where  $b_{11}$  represents the average bulk density at the smallest compaction soil water content,  $b_{13}$  is a relative increment of the density per unit of water content, and  $b_{12}$  is a scaling parameter.

The second phase is connected with approaching the maximum pore filling with water. This extreme condition is described with the function

$$p_d = \frac{100p_s}{100 + xp_s}.$$

Where  $p_d$  is the maximum soil bulk density at water content  $x$  and the specific density of soil  $p_s$  (in our case  $p_s$  was equal to  $2.61 \text{ g/cm}^3$ ). In result, the second phase was modeled simply by linear function of  $p_d$  i.e. in the form

$$f_2(x, b_2) = b_{21} + b_{22}p_d = b_{21} + b_{22} \frac{100p_s}{100 + xp_s}$$

Where  $b_{21}$  represent a constant shift and  $b_{22}$  is the rate of arriving to the maximum bulk density.

The phases were joined by a transition function of the form

$$h(x, g) = 1 - \frac{1}{1 + e^{\frac{x}{g^2}}} \quad (15)$$

It can be easily noted that the smaller values of  $g$  causes the faster change of  $h$  in the neighborhood of zero. Actually,  $h$  is a shifted and rescaled hyperbolic tangent.

The function (15)  $h$  fulfills the condition (9), when  $g^2 < \frac{1}{b_{13}}$ , and the condition (11), when  $b_{13} > 0$ . The other conditions are satisfied without any restriction. Since the inequality  $g^2 < \frac{1}{b_{13}}$  implies  $b_{13} > 0$ , the only restraint is the former one. It should be controlled during estimation process.

According to the method presented in Section 4, the domains of individual phases as well the estimates of unknown parameters were determined by the sequential procedure. Selected results are given in Table 1 and Table 2.

The optimal results for both phases were obtained coincidentally in the same step 31. comparing the domains of phases  $f_1$  and  $f_2$ , it follows that the transition point belong to the interval  $[12.63; 13.44]$ . In the second stage this interval was searched with the increment 0.01.

In the first case, only the parameter  $g$  and the transition point  $a$  were estimated. The following results were obtained:

$$R^2 = 0.856;$$

$$\hat{a} = 12.63;$$

$$\hat{g} = 0.573;$$

$$\hat{b}_{11} = 1.659; \quad \hat{b}_{12} = -6.336; \quad \hat{b}_{13} = 0.368$$

$$\text{and} \quad \hat{b}_{21} = -0.550, \quad \hat{b}_{22} = 1.248$$

Table 1  
Estimation of the first phase

Step	$b_{11}$	$b_{12}$	$b_{13}$	$R^2$	Domain
5	-21.677	3.152	0.0002	0.052	[5.51; 19.62]
10	-37.740	3.672	0.0004	0.373	[5.51; 18.41]
20	1.377	-1.590	0.0600	0.657	[5.51; 13.84]
25	1.582	-3.328	0.1500	0.683	[5.51; 14.81]
30	1.647	-5.445	0.3000	0.730	[5.51; 13.91]

Table 2  
Estimation of the second phase

Step	$b_{21}$	$b_{22}$	$R^2$	Domain
5	1.653	0.050	0.007	[5.63; 21.17]
10	1.435	0.166	0.052	[6.05; 21.17]
20	0.766	0.526	0.303	[9.03; 21.17]
25	0.203	0.834	0.648	[10.81; 21.17]
30	-0.418	1.175	0.929	[10.95; 21.17]
<b>31</b>	<b>-0.550</b>	<b>1.248</b>	<b>0.966</b>	<b>[12.63; 21.17]</b>
32	-0.553	1.249	0.965	[12.78; 21.17]
33	-0.560	1.253	0.962	[12.99; 21.17]

In the second case, all parameters  $b_{ij}$  of regression (8) were re-estimated, while estimating  $g$  and searching the transition point. The results are as follows:

$$R^2 = 0.858;$$

$$\hat{a} = 12.65;$$

$$\hat{g} = 0.221;$$

$$\hat{b}_{11} = 1.676; \quad \hat{b}_{12} = -9.07; \quad \hat{b}_{13} = 0.599$$

$$\text{and} \quad \hat{b}_{21} = -0.551, \quad \hat{b}_{22} = 1.248$$

Observe that in both cases the condition  $g^2 < \frac{1}{b_{13}}$  is satisfied.

Moreover, observe that the estimates obtained in the first case, except parameters  $b_{12}$  and  $b_{13}$ , coincide with those arrived from second one. Thus, in both cases the parameters have retained their physical interpretation and each of the resulting regressions describes similarly the investigated process.

We have also estimated the function (8) by searching the transition point throughout the whole range of variability for independent variable. The following results were obtained:

$$R^2 = 0.863;$$

$$\hat{a} = 121.64;$$

$$\hat{g} = 0.625;$$

$$\hat{b}_{11} = -6.967; \quad \hat{b}_{12} = 2.151; \quad \hat{b}_{13} = 0.000944$$

$$\text{and} \quad \hat{b}_{21} = -0.584, \quad \hat{b}_{22} = 1.267$$

It can be seen that the estimate of the phase one differs significantly from the corresponding phases established earlier. On the other hand, we can note quite well fit of the joined regression to the observed points. It is entirely due to the transition function. Therefore, confining the search of the transition point to the shortest possible interval, is advantageous for the final interpretation of the regression.

### References

- [1] Hinkley, D. V. Inference about the intersection in two-phase regression, 1969
- [2] Lerman, P. M. Fitting segmented regression models by grid search, 1980
- [3] Serber, G. A. and C. J. Wild. Nonlinear regression, 1989
- [4] Sprent, P. Some hypotheses concerning two-phase regression lines, 1961
- [5] Watts, D. G. and D. W. Bacon. Using an hyperbola as a transition model to fit two-regime straight-line data, 1974

### Author:

**Stoyan Ivanov Nenkov**

Varna Free University, Bulgaria

[mrnenkov@abv.bg](mailto:mrnenkov@abv.bg)

## DECISION POWER INDICES IN WEIGHTED VOTING GAME

Zdravko Slavov, Christina Evans

**Abstract.** In this paper we consider some properties of decision power indices in weighted voting games. We discuss two most widely used measures of voting power in weighted voting games - the Shapley-Shubik power index and the Banzhaf power index. The Shapley-Shubik power index in a voting situation depends on the number of orderings in which each player is pivotal. The Banzhaf power index depends on the number of ways in which each player can affect a negative swing.

**Keywords:** weighted voting game, power index, winning coalition, Shapley-Shubik, Banzhaf.

### 1. Introduction

The notion of a simple game was introduced by John von Neumann and Oscar Morgenstern in their monumental book "Theory of Games and Economic Behavior" in 1944 [10]. This game is a conflict in which the only objective is winning and the only rule is an algorithm to decide which coalitions are winning.

Definition 1. A simple game is a pair  $(N, v)$  where  $N = \{1, 2, \dots, n\}$  is the set of players,  $n = |N| \geq 2$ , and  $v: 2^N \rightarrow \{0, 1\}$  is the characteristic function which satisfies the following three conditions:

- (1)  $v(\emptyset) = 0$ .
- (2)  $v(N) = 1$ .
- (3) if  $S$  and  $T$  are coalitions of players in  $N$  such that  $S \subset T$ , then  $v(S) \leq v(T)$ .

The characteristic function  $v$  for a coalition  $S$  indicates the value of  $S$ . For each coalition  $S \subset N$  there is either  $v(S) = 0$  or  $v(S) = 1$ .

Definition 2. For a coalition  $S \subset N$ , we define that coalition  $S$  is winning if and only if  $v(S) = 1$ , coalition  $S$  is losing if and only if  $v(S) = 0$ , and the losing coalition  $S$  is blocking if and only if  $N \setminus S$  is a losing coalition too. The collection of all winning coalitions is denoted by  $W$ .

By definition  $N \in W$ ; therefore,  $W$  is nonempty.

Definition 3. Let us label the players by  $1, 2, \dots, n$  and let  $S \in W$ ,  $S$  is called a minimal winning coalition if and only if  $S \setminus \{i\}$  is a losing coalition for all  $i \in S$ . The collection of all minimal winning coalitions is denoted by  $MW$ .

It is easy to prove that  $MW \subset W$  is nonempty.

Definition 4. A player who does not belong to any minimal winning coalition is called a dummy, i.e. player  $i \in N$  is a dummy if and only if



$i \notin S$  for all  $S \in MW$ . A player who belongs to all minimal coalitions is called a veto player, i.e. a player  $i \in N$  has capacity to veto if and only if  $i \in S$  for all  $S \in MW$ . A player  $i \in N$  is a dictator if and only if  $\{i\}$  is a winning coalition, i.e.  $\{i\} \in MW$ .

## 2. Weighted Voting Games

We will consider a special class of simple games called weighted voting games with dichotomous voting rule - acceptance or rejection. The basic formal framework of this study is as follows. The symbol  $[q; w_1, w_2, \dots, w_n]$  we will be used, where  $q$  and  $w_1, w_2, \dots, w_n$  are nonnegative integer numbers such that  $w_i \leq q \leq \sum_{k=1}^n w_k = t$  for all  $i \in N$ . By convention, we take  $w_i \geq w_j$  if  $i < j$ . Here we have the following properties:

- (1)  $1 \leq q \leq t$ .
- (2)  $n = |N| \geq 2$  is the number of players.
- (3)  $w_i \geq 0$  is the number of votes of player  $i \in N$  and  $w_1 \geq 1$ .
- (4)  $q$  is the needed quota so that a coalition can win.

The symbol  $[q; w_1, w_2, \dots, w_n]$  represents the simple game  $(N, v)$  defined by

$$v(S) = \begin{cases} 1, & \sum_{k \in S} w_k \geq q \\ 0, & \sum_{k \in S} w_k < q \end{cases},$$

where  $S \subset N$ .

**Remark 1.** A simple game  $(N, v)$  is called proper if and only if  $v(S) + v(N \setminus S) \leq 1$  for all  $S \subset N$ . In a proper game,  $S$  and  $N \setminus S$  cannot both be winning coalitions. It is easy to show that if  $2q > t$ , then the weighted voting game  $(N, v)$  is proper. In an improper game there will be at least one pair of non-intersection winning coalitions. If  $v(S) + v(N \setminus S) = 1$  for all  $S \subset N$ , then the game is said to be decisive.

**Example 1.** The voting method of the Security Council of the United Nations, formed by 5 permanent and 10 temporary members, is the game in which each one of the permanent members has 7 votes and each one of the temporary members has only one vote, the established quota is 39 votes, the total is 45 votes. We observe that any coalition which does not include all of the 5 permanent members has at most  $4 \times 7 + 10 = 38$  votes, which is an inferior number to the fixed quota. As a result this

coalition will not be winning. Hence, each one of the permanent members has capacity to veto any proposal. For more information see [5] and [9].

Example 2. The Bulgarian Parliament has 240 seats and uses two different rules: a simple majority by quota 121 (more of 50%) and a qualified majority by quota 161 (more of  $\frac{2}{3}$ ). The Parliament of Finland has 200 seats and uses three different rules: a simple majority by quota 101 (more of 50%), a qualified majority by quota 134 (more of  $\frac{2}{3}$ ), and in some special case by quota 167 (more of  $\frac{5}{6}$ ) [5].

Example 3. The U. S. Congress has a nonvoting delegate who represents the District of Columbia; therefore, this delegate is a dummy.

Example 4. Note that in principle a player can be assigned weight zero, but in practice this player would be silly, because it would be a dummy. However, a player having positive weight can also be a dummy. It is known that Luxembourg as a member of the European Union Council of Ministers in the period 1958-1973 had weight one but was a dummy. Now, we call this situation the voting paradox of Luxembourg [8].

For  $m \geq 2$ , let  $(N, v_1), (N, v_2), \dots, (N, v_m)$  be simple games. We can now study the two new simple games  $(N, v_1 \wedge v_2 \wedge \dots \wedge v_m)$  and  $(N, v_1 \vee v_2 \vee \dots \vee v_m)$  defined by

$$(v_1 \wedge v_2 \wedge \dots \wedge v_m)(S) = \min\{v_k(S) : 1 \leq k \leq m\},$$

$$(v_1 \vee v_2 \vee \dots \vee v_m)(S) = \max\{v_k(S) : 1 \leq k \leq m\}.$$

A weighted  $m$ -majority game is a simple game  $(N, v_1 \wedge v_2 \wedge \dots \wedge v_m)$  where the games  $(N, v_k)$  for  $k=1,2,\dots,m$  are the weighted voting games represented by  $[q^k; w_1^k, w_2^k, \dots, w_n^k]$ . Hence, the characteristic function is given by  $(v_1 \wedge v_2 \wedge \dots \wedge v_m)(S) = 1$  if  $\sum_{i \in S} w_i^k \geq q^k$  for all  $k=1,2,\dots,m$  and  $(v_1 \wedge v_2 \wedge \dots \wedge v_m)(S) = 0$  otherwise (if  $\sum_{i \in S} w_i^k < q^k$  for some  $k=1,2,\dots,m$ ).

If  $m=2$  or  $m=3$ , then we obtain a weighted double or triple majority game, respectively.

### 3. Decision Power Indices of the Players

Weighted voting games use mathematical models which are used to analyze the distribution of decision power of the players. This analysis

of power is central in political science. In general, it is difficult to define the idea of power, but for the special case of voting power mathematical power indices have been used. There are two most widely used measures of voting power in weighted voting games - the Shapley-Shubik power index and the Banzhaf power index. The Shapley-Shubik power index in a voting situation depends on the number of orderings in which each player is pivotal. The Banzhaf power index depends on the number of ways in which each player can affect a negative swing.

### 3.1. The Shapley-Shubik Power Index

The Shapley-Shubik power index was introduced by the mathematician Lloyd Shapley and the economist Martin Shubik in 1954 [7]. For player  $i \in N$  this index is defined by

$$f_i = \sum_{S \subseteq W, S \cup \{i\} \in W} \frac{s!(n-s-1)!}{n!},$$

where  $s = |S|$ . If we assume that all  $n!$  orderings are equiprobable, then  $f_i$  is the probability of player  $i$  being pivotal in a winning coalition, that is,  $S \cup \{i\}$  is a winning and  $S$  is a losing coalition.

Remark 2. In classical theory, a negative swing for player  $i \in N$  is a pair of coalitions  $(S \cup \{i\}, S)$  such that  $S \cup \{i\}$  is winning and  $S$  is losing, i.e.  $i \notin S$ ,  $v(S \cup \{i\}) = 1$  and  $v(S) = 0$ . It is easy to show that  $v(S) - v(S \setminus \{i\})$  is always either zero or one for all  $S \subset N$  and all  $i \in N$ . If  $S \subset N$  and  $i \notin S$ , then  $v(S) - v(S \setminus \{i\}) = 0$ . If  $S \subset N$  and  $i \in S$ , then  $v(S) - v(S \setminus \{i\}) = 1$  (when  $S$  is winning and  $S \setminus \{i\}$  is losing) or  $v(S) - v(S \setminus \{i\}) = 0$  (when  $S$  and  $S \setminus \{i\}$  are both winning or  $S$  and  $S \setminus \{i\}$  are both losing).

From Remark 2 it follows that

$$f_i = \sum_{S \subset N} \frac{(s-1)!(n-s)!}{n!} (v(S) - v(S \setminus \{i\})) = \sum_{S \subset N: i \in S} \frac{(s-1)!(n-s)!}{n!} (v(S) - v(S \setminus \{i\})).$$

For each player  $i \in N$  we also obtain

$$f_i = \sum_{j=0}^{n-1} \frac{j!(n-j-1)!}{n!} d_j^i,$$

where each  $d_j^i$  is the number of negative swings of player  $i$  in a coalition of size  $j$ .

The Shapley-Shubik index is the vector  $f = (f_1, f_2, \dots, f_n)$  and it has the normalization property, i.e.  $\sum_{i=1}^n f_i = 1$ .

### 3.2. The Banzhaf Power Index

The Banzhaf power index was introduced by the american jurist and law professor John Banzhaf III in 1965 [3]. The absolute Banzhaf index concerns the number of times each player  $i \in N$  could change a coalition from losing to winning and it requires that we know the number of negative swings for each player  $i$ . For each player  $i \in N$ , the absolute Banzhaf index is denoted by  $h_i$  and it equals the number of negative swings for this player.

The normalized Banzhaf power index is the vector  $b = (b_1, b_2, \dots, b_n)$ , given by

$$b_i = \frac{h_i}{\sum_{k=1}^n h_k} \text{ for } i=1,2,\dots,n.$$

Remark 3. It is easy to show that  $h_i = \sum_{S \subset N} (v(S) - v(S \setminus \{i\})) = \sum_{S \subset N: i \in S} (v(S) - v(S \setminus \{i\}))$  for all  $i=1,2,\dots,n$ . See also Remark 2. The Banzhaf index is similar of the Penrose-Banzhaf (or Banzhaf-Coleman) index which is defined by  $b_i = \sum_{S \subset N} \frac{v(S) - v(S \setminus \{i\})}{2^{n-1}} = \frac{h_i}{2^{n-1}}$  for  $i=1,2,\dots,n$ .

Remark 4. Let  $S \subset N$  be a minimal winning coalition and  $i \in N$ . It is easy to show that  $v(S) - v(S \setminus \{i\}) = 1$  for all  $i \in S$  and  $v(S) - v(S \setminus \{i\}) = 0$  for all  $i \notin S$ .

### 4. The Power Distribution of Decision Power in EU

The power of a country in a supranational organization is a numerical measure of its capacity to decide the approval of a motion. This decisive character is measured by calculating the number of times that the vote of a country converts to a coalition that does not reach the quota to make decisions into a winning coalition.

The most important institutions of the European Union are as follows:

- (1) The European Parliament which represents the citizens and it is elected directly by them;
- (2) The Council of Ministers which represents the individual member states;

(3) The European Commission which is meant to uphold the interests of the Union as a whole.

In principle, the European Commission proposes and applies policies and laws. The European Parliament and the Council of Ministers adopt these policies and laws.

The power indices have received increasing attention in political science, especially in the field of European Union politics. They are frequently used for investigating their present distribution of voting power among EU member states in the Council of Ministers and the European Parliament.

The focus of this section will be on the Council of Ministers. We will present some results concerning to the Shapley-Shubik and the normalized Banzhaf power indices in the decision process in the European Union.

#### 4.1. The Nice European Council in EU-27

The Council of Ministers of EU, shortly CM, represents the national governments of the member states. The Council uses a voting system of qualified majority to pass new legislation. The Nice European Council on December 2000 established two decision rules. These rules are contained in the “Declaration on the enlargement of the European Union” and the “Declaration on the qualified majority threshold and the number of votes for a blocking minority in an enlarged Union”.

The first decision rule of CM, shortly Nice1. is the weighted triple majority game  $v_1 \wedge v_2 \wedge v_3$ , where the three weighted voting games corresponding to votes, countries and population, are the following:

$$v_1 = [255; 29, 29, 29, 29, 27, 27, 14, 13, 12, 12, 12, 12, 10, 10, 10, 7, 7, 7, 7, 4, 4, 4, 4, 3];$$

$$v_2 = [14; 1, 1];$$

$$v_3 = [620; 170, 123, 221, 208, 280, 47, 33, 22, 21, 21, 21, 18, 17, 17, 11, 11, 18, 8, 5, 4, 3, 2, 1].$$

In this game the players are: {Germany, UK, France, Italy, Spain, Poland, Romania, Netherlands, Greece, Czech, Belgium, Hungary, Portugal, Sweden, Bulgaria, Austria, Slovakia, Denmark, Finland, Ireland, Lithuania, Latvia, Slovenia, Estonia, Cyprus, Luxembourg, Malta}.

Note that in a game with 27 players the total number of coalitions which can be formed is  $2^{27} = 134,217,728$  and the number of nonempty coalitions is 134,217,727.

From the weighted triple majority game  $v_1 \wedge v_2 \wedge v_3$  it follows that a voting will be favorable if it counts the support of 14 countries with at least 255 votes, and at least 62% of the population. See Tables 1 and 2.

Table 1. The Power Distribution under the Decision rule Nice1 [4]

Country	Votes	Shapley-Shubik	normalized Banzhaf
Germany	29	0,0871	0,0778
UK	29	0,0870	0,0778
France	29	0,0870	0,0778
Italy	29	0,0870	0,0778
Spain	27	0,0799	0,0742
Poland	27	0,0799	0,0742
Romania	14	0,0399	0,0426
Netherlands	13	0,0368	0,0397
Greece	12	0,0340	0,0368
Czech	12	0,0340	0,0368
Belgium	12	0,0340	0,0368
Hungary	12	0,0340	0,0368
Portugal	12	0,0340	0,0368
Sweden	10	0,0281	0,0309
Bulgaria	10	0,0281	0,0309
Austria	10	0,0281	0,0309
Slovakia	7	0,0196	0,0218
Denmark	7	0,0196	0,0218
Finland	7	0,0196	0,0218
Ireland	7	0,0196	0,0218
Lithuania	7	0,0196	0,0218
Latvia	4	0,0110	0,0125
Slovenia	4	0,0110	0,0125
Estonia	4	0,0110	0,0125
Cyprus	4	0,0110	0,0125
Luxembourg	4	0,0110	0,0125
Malta	3	0,0082	0,0094

The second decision rule in CM, shortly Nice2. is the weighted triple majority game  $v_1 \wedge v_2' \wedge v_3$ , where the weighted voting game  $v_2'$  consists of a qualified majority of  $\frac{2}{3}$  of the countries; therefore, it is as follows:

$$v'_2 = [18; 1, 1].$$

For more information about the decision rule Nice2 and the power distribution see Table 2.

Table 2. The Power Distribution under the Decision rule Nice2 [4]

Country	Share of Population	Shapley-Shubik	normalized Banzhaf
Germany	0,170	0,0837	0,0665
UK	0,123	0,0836	0,0665
France	0,122	0,0836	0,0665
Italy	0,120	0,0836	0,0665
Spain	0,082	0,0767	0,0631
Poland	0,080	0,0767	0,0631
Romania	0,047	0,0394	0,0407
Netherlands	0,033	0,0365	0,0386
Greece	0,022	0,0340	0,0366
Czech	0,021	0,0340	0,0366
Belgium	0,021	0,0340	0,0366
Hungary	0,021	0,0340	0,0366
Portugal	0,021	0,0340	0,0366
Sweden	0,018	0,0286	0,0325
Bulgaria	0,017	0,0286	0,0325
Austria	0,017	0,0286	0,0325
Slovakia	0,011	0,0208	0,0263
Denmark	0,011	0,0208	0,0263
Finland	0,011	0,0208	0,0263
Ireland	0,008	0,0208	0,0263
Lithuania	0,008	0,0208	0,0263
Latvia	0,005	0,0131	0,0198
Slovenia	0,004	0,0131	0,0198
Estonia	0,003	0,0131	0,0198
Cyprus	0,002	0,0131	0,0198
Luxembourg	0,001	0,0131	0,0198
Malta	0,001	0,0106	0,0177

Table 3. The Power Distribution under the Decision rules Nice1  
from the Bulgarian point of view

Country	Nice1	
	Shapley-Shubik	normalized Banzhaf
Germany	3,10	2,52
UK	3,10	2,52
France	3,10	2,52
Italy	3,10	2,52
Spain	2,84	2,40
Poland	2,84	2,40
Romania	1,42	1,38
Netherlands	1,31	1,28
Greece	1,21	1,19
Czech	1,21	1,19
Belgium	1,21	1,19
Hungary	1,21	1,19
Portugal	1,21	1,19
Sweden	1,00	1,00
Bulgaria	1,00	1,00
Austria	1,00	1,00
Slovakia	0,70	0,71
Denmark	0,70	0,71
Finland	0,70	0,71
Ireland	0,70	0,71
Lithuania	0,70	0,71
Latvia	0,39	0,40
Slovenia	0,39	0,40
Estonia	0,39	0,40
Cyprus	0,39	0,40
Luxembourg	0,39	0,40
Malta	0,29	0,30

Table 4. The Power Distribution under the Decision rules Nice2  
from the Bulgarian point of view

Country	Nice2	
	Shapley-Shubik	normalized Banzhaf
Germany	2,93	2,05
UK	2,92	2,05
France	2,92	2,05
Italy	2,92	2,05





Furthermore, the minimum number of countries to block a proposal is 4 and abstentions are not counted.

The Treaty of Lisbon specifies that until 31 October 2014 the Nice decision rules will be used.

In this game the players are: {Germany, France, UK, Italy, Spain, Poland, Romania, Netherlands, Greece, Portugal, Belgium, Czech, Hungary, Sweden, Austria, Bulgaria, Denmark, Slovakia, Finland, Ireland, Lithuania, Latvia, Slovenia, Estonia, Cyprus, Luxembourg, Malta}. Here there is a new order of the players.

The power distribution of the countries under the European Constitution rule is showed in Table 5.

Table 5. The Power Distribution under the Constitution rule [6]

Country	Shapley-Shubik	normalized Banzhaf
Germany	0,1627	0,1187
France	0,1088	0,0874
UK	0,1082	0,0869
Italy	0,1041	0,0844
Spain	0,0737	0,0637
Poland	0,0681	0,0589
Romania	0,0421	0,0422
Netherlands	0,0326	0,0350
Greece	0,0242	0,0288
Portugal	0,0232	0,0280
Belgium	0,0232	0,0280
Czech	0,0229	0,0278
Hungary	0,0228	0,0277
Sweden	0,0209	0,0263
Austria	0,0195	0,0252
Bulgaria	0,0192	0,0249
Denmark	0,0152	0,0219
Slovakia	0,0152	0,0219
Finland	0,0150	0,0217
Ireland	0,0131	0,0202
Lithuania	0,0123	0,0196
Latvia	0,0105	0,0182
Slovenia	0,0100	0,0178
Estonia	0,0090	0,0170
Cyprus	0,0080	0,0162
Luxembourg	0,0076	0,0159
Malta	0,0075	0,0158

Table 6. The Power Distribution under the Constitution rule  
from the Bulgarian point of view

Country	Shapley-Shubik	normalized Banzhaf
Germany	8,47	4,77
France	5,67	3,71
UK	5,66	3,49
Italy	5,42	3,39
Spain	3,84	2,52
Poland	3,57	2,37
Romania	2,19	1,69
Netherlands	1,70	1,41
Greece	1,26	1,16
Portugal	1,21	1,12
Belgium	1,21	1,12
Czech	1,19	1,12
Hungary	1,19	1,11
Sweden	1,09	1,06
Austria	1,02	1,01
Bulgaria	1,00	1,00
Denmark	0,79	0,88
Slovakia	0,79	0,88
Finland	0,78	0,87
Ireland	0,68	0,81
Lithuania	0,64	0,79
Latvia	0,55	0,73
Slovenia	0,52	0,71
Estonia	0,47	0,68
Cyprus	0,42	0,65
Luxembourg	0,40	0,64
Malta	0,39	0,63

The numbers of negative swings in the double majority game  $v_2'' \wedge v_3''$ , shortly Game2, and in the European Constitution game  $(v_2'' \wedge v_3'') \vee v_4$ , shortly Game3, are showed in Table 7.

Table 7. Numbers of negative swings for EU-27 [1]

Country	Game2	Game3	Differences
Germany	13,708,647	13,708,639	-8
France	10,090,873	10,090,873	0
UK	10,039,263	10,039,263	0
Italy	9,743,361	9,743,561	0
Spain	7,362,225	7,362,229	+4
Poland	6,800,035	6,800,039	+4
Romania	4,871,079	4,871,089	+10
Netherlands	4,048,155	4,048,165	+10
Greece	3,326,413	3,326,423	+10
Portugal	3,237,845	3,237,855	+10
Belgium	3,231,183	3,231,193	+10
Czech	3,210,689	3,210,699	+10
Hungary	3,196,981	3,196,991	+10
Sweden	3,033,941	3,033,951	+10
Austria	2,911,521	2,911,531	+10
Bulgaria	2,877,541	2,877,551	+10
Denmark	2,530,351	2,530,361	+10
Slovakia	2,530,351	2,530,361	+10
Finland	2,509,871	2,509,881	+10
Ireland	2,332,655	2,332,665	+10
Lithuania	2,264,209	2,264,219	+10
Latvia	2,099,685	2,099,695	+10
Slovenia	2,051,511	2,051,521	+10
Estonia	1,962,711	1,962,721	+10
Cyprus	1,873,045	1,873,055	+10
Luxembourg	1,831,725	1,831,735	+10
Malta	1,824,817	1,824,827	+10

### References

- [1] Algaba E., J. Bilbao, J. Fernandez, The distribution of power in the European Constitution, *European Journal of Operational Research* 176 (2007), 1752-1766.
- [2] Alonso-Meijide J., J. Bilbao, Casas-Mendes B., J. Fernandez, Weighted multiple majority games with unions: Generating functions and applications to the European Union, *European Journal of Operational Research* 198 (2009), 530-544.
- [3] Banzhaf III J., Weighted voting doesn't work: A mathematical analysis, *Rutgers Law Review* 19 (1965), 317-343.

- [4] Bilbao J., J. Fernandez, N. Jimenez, J. Lopez, Voting power in the European Union enlargement, *European Journal of Operational Research* 143 (2002), 181-196.
- [5] Laruelle A., F. Valenciano, *Voting and Collective Decision-Making: Bargaining and Power*, Cambridge University Press, 2008.
- [6] Paterson I., Voting Power Derives from the Poll Distribution: Shedding Light on Contentious Issues of Weighted Votes and the Constitutional Treaty, ENEPRI, working paper N: 50, February 2007.
- [7] Shapley L., M. Shubik, A method for evaluating the distribution of power in a committee system, *American Political Science Review* 48 (1954), 787-792.
- [8] Slavov Z., C. Evans, On the Voting Paradox of Luxembourg and Decision Power Indices, *Mathematics and Education in Mathematics* 43 (2014), 138-144.
- [9] Taylor A., A. Pacelli, *Mathematics and Politics: Strategy, Voting, Power and Proof*. Stringer, 2008.
- [10] Von Newman J., O. Morgenstern, *Theory of Game and Economic Behavior*, Princeton University Press, 1944.

**Authors:**

**prof. Zdravko Dimitrov Slavov**, PhD  
Varna Free University, Bulgaria  
[slavovibz@yahoo.com](mailto:slavovibz@yahoo.com)

**Christina Slavova Evans**

The George Washington University, Washington DC, USA  
[evans.christina.s@gmail.com](mailto:evans.christina.s@gmail.com)

## DISTANCE, SIMILARITY AND DISSIMILARITY MEASURES IN STATISTICAL ANALYSIS OF BINARY DATA

Zdravko Slavov, Desislava Slavova

**Abstract.** *In the present paper we discuss the concept of distance, similarity and dissimilarity measures in statistical analysis of binary data. We show that the key ideas of similarity and dissimilarity measures of binary data is an application the concepts of the distance measures and the Pearson correlation coefficient. The binary similarity and dissimilarity measures play a fundamental role in statistics, computer science, biology, ethnology, taxonomy, image retrieval, geology and chemistry.*

**Keywords:** *binary data, measure, distance, correlation, similarity, dissimilarity*

### 1. Introduction

A binary attribute is an attribute that has exactly two possible values such as: “1” and “0”, “yes” and “no”, “true” and “false”, “male” and “female”, “large” and “small”, and so on. The binary variables can be divided into two types: symmetric and asymmetric. In a symmetric binary variable we always have that the two values are equivalent. But in asymmetric variable we have that one of its values carries more important information than the other.

It is well-known that binary similarity and dissimilarity measures play a fundamental role in statistical analysis of data in classification and clustering [2] [3] [4] [5] [6] [12] [13] [14]. Numerous binary similarity measures have been proposed in various fields. For example, the Jaccard similarity measure was used for clustering ecological species and Forbes proposed a coefficient for clustering ecologically related species [7]. The binary similarity measures were subsequently applied in biology, ethnology, taxonomy, image retrieval, geology and chemistry [8] [10] [11] [15] [17]. Recently, they have been actively used as fingerprints, iris images and handwritten character recognitions [1] [18].

The binary similarity measures have been described in the literature. In [8], we see a collection of 43 similarity measures, and 20 of them were used for cluster analysis on fungi data to produce five clusters of related coefficients. In [9], one can read about the best 8 binary similarity measures that researchers use in ecology. In [18], there are 13 similarity measures for binary fingerprint code. In [2], we see 76 definition of binary similarity and dissimilarity measures.

## 2. Distance measure

Let us denote a distance measure between two points  $x, y \in R^m$  by  $d(x, y) \in R_+$ ,  $m \geq 1$ . Recall that for the function distance measure  $d : R^m \times R^m \rightarrow R_+$  (or distance metric) the following statements are true:

(1)  $d(x, y) \geq 0$  for all  $x, y \in R^m$ .

Distance is not negative between two points.

(2)  $d(x, y) = 0$  if and only if  $x = y$  for every two points  $x, y \in R^m$ .

Distance is positive between two different points; it is zero precisely from a point to itself.

(3)  $d(x, y) = d(y, x)$  for all  $x, y \in R^m$ .

The distance between  $x$  and  $y$  is the same in either direction.

(4)  $d(x, y) + d(y, z) \geq d(x, z)$  for all  $x, y, z \in R^m$ .

The distance measure between two points is the shortest distance along any path. This condition is called the triangle inequality.

In a general form, the distance measure between two points

$x, y \in R^m$  is defined as  $d(x, y) = \left( \sum_{k=1}^m |x_k - y_k|^r \right)^{1/r}$  where  $r \geq 1$  is a real parameter. It is also called a Minkowski distance or  $L_r$ -norm. The real parameter  $r$  need not be an integer, but it cannot be less than 1, because otherwise the triangle inequality does not hold.

The distance measure  $d(x, y) = \left( \sum_{k=1}^m w_k |x_k - y_k|^r \right)^{1/r}$  is called a weighted Minkowski distance for the real numbers  $w_k > 0$  for  $k = 1, 2, \dots, m$ .

The following is a list of common distances for specific values of the parameter  $r$ .

(1) For  $r = 1$  it is called the absolute distance (or city block, or Block distance, or Manhattan distance, or taxicab, or  $L_1$ -norm).

This is  $d(x, y) = \sum_{k=1}^m |x_k - y_k|$ .

(2) For  $r = 2$  it is called the Euclidean distance (or  $L_2$ -norm).

This is  $d(x, y) = \sqrt{\sum_{k=1}^m (x_k - y_k)^2}$ .

(3) For  $r \rightarrow \infty$  it is called the Ghebyshev distance (or maximum distance, or  $L_\infty$ -norm).

This is the maximum difference between any component of the points or  $d(x, y) = \max_{1 \leq k \leq m} |x_k - y_k|$ .

### 3. Pearson correlation coefficient

The Pearson correlation coefficient for continuous data can be

written as  $r(p, q) = \frac{\sum_{k=1}^m (p_k - \bar{p})(q_k - \bar{q})}{\sqrt{\sum_{k=1}^m (p_k - \bar{p})^2 \cdot \sum_{k=1}^m (q_k - \bar{q})^2}} = \frac{\text{cov}(p, q)}{s_p s_q}$ , where

$$p, q \in R^m, \quad \bar{p} = \frac{1}{m} \sum_{k=1}^m p_k, \quad \bar{q} = \frac{1}{m} \sum_{k=1}^m q_k,$$

$\text{cov}(p, q) = \frac{1}{m} \sum_{k=1}^m (p_k - \bar{p})(q_k - \bar{q})$ ,  $s_p > 0$  and  $s_q > 0$ . This is also the cosine of the angle between the following two vectors:  $(p_1 - \bar{p}, p_2 - \bar{p}, \dots, p_m - \bar{p})$  and  $(q_1 - \bar{q}, q_2 - \bar{q}, \dots, q_m - \bar{q})$ . Historically, this correlation coefficient was first discovered by Bravais (1846) and was later shown by Pearson to be the best possible correlation coefficient.

The Pearson correlation coefficient can also be written as

$$r = \frac{1}{m} \sum_{k=1}^m \left( \frac{p_k - \bar{p}}{s_p} \right) \left( \frac{q_k - \bar{q}}{s_q} \right). \text{ The coefficient varies between } -1 \text{ and } +1.$$

The case when  $r = +1$ , called perfect positive correlation, occurs when  $p$  and  $q$  perfectly coincide, and the case when  $r = -1$ , called perfect negative correlation, occurs when  $p$  and  $-q$  perfectly coincide. Under perfect positive or negative correlation we have that  $\frac{p_k - \bar{p}}{s_p} = \pm \frac{q_k - \bar{q}}{s_q}$

for  $k = 1, 2, \dots, m$  or  $q = \pm \frac{s_q}{s_p} (p - \bar{p}) + \bar{q}$ , showing that the corresponding

$p$  and  $q$  values are related linearly. When  $r$  is not equal to  $+1$  or  $-1$ , the line best fitting the corresponding values of  $p$  and  $q$  is obtained from

$$q' = r \frac{s_q}{s_p} (p - \bar{p}) + \bar{q}. \text{ Therefore, the correlation coefficient } r \text{ can be}$$

considered to be a coefficient of the linear relationship between the corresponding values of  $p$  and  $q$ .



#### 4. Similarity and dissimilarity measures

For function similarity measures  $S : R^m \times R^m \rightarrow [0;1]$  and  $\tilde{S} : R^m \times R^m \rightarrow [-1;1]$  the following statements are true:

- (1)  $S(p, p) = 1$  and  $\tilde{S}(p, p) = 1$  for all  $p \in R^m$  (reflectivity).
- (2)  $S(p, q) < 1$  and  $\tilde{S}(p, q) < 1$  for every  $p, q \in R^m$  such that  $p \neq q$ .
- (3)  $S(p, q) = S(q, p)$  and  $\tilde{S}(p, q) = \tilde{S}(q, p)$  for all  $p, q \in R^m$  (symmetry).

Note that the similarity measure  $\tilde{S}$  is usually called a correlation coefficient and we have that  $S(p, q) = |\tilde{S}(p, q)|$  or  $S(p, q) = \tilde{S}^2(p, q)$  for all  $p, q \in R^m$ . It is known that for each correlation coefficient we also have  $\tilde{S}(p, -q) = -\tilde{S}(p, q)$ ; therefore,  $\tilde{S}(p, -p) = -1$  for all  $p \in R^m$ .

It is easy to show that if  $d$  is a distance metric, then  $S = \frac{1}{1+d}$  is a similarity measure. In general,  $S_1 = \frac{1}{a+d}$  is a similarity measure for every real number  $a > 0$ .

Let the function similarity measure  $S$  be given and  $g : [0,1] \rightarrow [0,1]$  be a real increasing continuous transformation such that  $g(0) = 0$  and  $g(1) = 1$ . In this case the function  $g \circ S$  is a similarity measure too. For example, if  $S$  is a similarity measure, then  $S^2$  and  $S^3$  are also similarity measures.

There are axioms for dissimilarity analogous to those for similarity. We usually require that a function measure  $D : R^m \times R^m \rightarrow [0;1]$  is referred to as a dissimilarity if it satisfies:

- (1)  $D(p, p) = 0$  for all  $p \in R^m$  (reflectivity).
- (2)  $D(p, q) > 0$  for every  $p, q \in R^m$  such that  $p \neq q$  (non-negativity).
- (3)  $D(p, q) = D(q, p)$  for all  $p, q \in R^m$  (symmetry).

Note that a similarity measure can easily be turned into a dissimilarity measure by one of the transformations  $D(p, q) = 1 - S(p, q)$  or  $D_1(p, q) = \sqrt{1 - S(p, q)}$  for all  $p, q \in R^m$ . So as a result of this concept we obtain two new dissimilarity measures

$$D(p, q) = 1 - S(p, q) = 1 - \frac{1}{1 + d(p, q)} = \frac{d(p, q)}{1 + d(p, q)} \quad \text{and}$$

$$D(p, q) = \sqrt{\frac{d(p, q)}{1 + d(p, q)}} \quad \text{for all } p, q \in R^m. \text{ In general, } D = \frac{d}{a + d} \text{ and}$$

$$D_1 = \sqrt{\frac{d}{a + d}} \text{ are dissimilarity measures for every real number } a > 0.$$

Let the function dissimilarity measure  $D$  be given and  $g : [0, 1] \rightarrow [0, 1]$  be a real increasing continuous transformation such that  $g(0) = 0$  and  $g(1) = 1$ . In this case the function  $g \circ D$  is a dissimilarity measure too. For example, if  $D$  is a similarity measure, then  $D^2$  and  $D^3$  are also similarity measures.

Another form of a function dissimilarity measure is  $D : R^m \times R^m \rightarrow R_+$ . Using the idea of the Minkowski distance metric we can define a Minkowski dissimilarity function by

$$D(p, q) = \left( \sum_{k=1}^m |p_k - q_k|^r \right)^{1/s} \text{ for all } p, q \in R^m, \text{ where } r \geq 1 \text{ and } s \geq 1 \text{ are}$$

real parameters. For example, when  $r = 2$  and  $s = 1$  we obtain the Squared Euclidean dissimilarity measure  $d(p, q) = \sum_{k=1}^m (p_k - q_k)^2$ . Let

$g : [0, 1] \rightarrow [0, 1]$  be a real increasing continuous transformation such that  $g(0) = 0$ . In this case the function  $g \circ D$  is a dissimilarity measure too.

For example, if  $D$  is a similarity measure, then  $D^2$  and  $D^3$  are also similarity measures.

## 5. Binary similarity measures

There are many similarity measures between binary vectors  $p$  and  $q$ ,  $p, q \in \{0, 1\}^m$  (or between dichotomous data). These measures are referred to as binary similarity measures or binary similarity coefficients, and they have values between 0 and 1 (or between -1 and 1). A value of 1 indicates that the two vectors are completely similar. A value of 0 indicates that the two vectors are not at all similar. There are many rationales for why one coefficient is better than another in specific instances.

The comparison of two binary vectors,  $p$  and  $q$ , leads to four quantities:

(1) Let  $m_{11}$  be equal to the number of positions where  $p$  is 1 and  $q$  is 1.

(2) Let  $m_{10}$  be equal to the number of positions where  $p$  is 1 and  $q$  is 0.

(3) Let  $m_{01}$  be equal to the number of positions where  $p$  is 0 and  $q$  is 1.

(4) Let  $m_{00}$  be equal to the number of positions where  $p$  is 0 and  $q$  is 0.

Each position must fall into one of these quantities, meaning that  $m_{11} + m_{10} + m_{01} + m_{00} = m$ .

Together, the four quantities  $m_{11}$ ,  $m_{10}$ ,  $m_{01}$  and  $m_{00}$  can be used to construct the  $2 \times 2$  contingency table:

$p \setminus q$	Value 1	Value 0	Total
Value 1	$m_{11}$	$m_{10}$	$p_1$
Value 0	$m_{01}$	$m_{00}$	$p_0$
Total	$q_1$	$q_0$	$m$

Similarity coefficients are defined as fractions and they have values between 0 and 1 or between -1 and 1.

Here the key idea is that the diagonal entries  $m_{11}$  and  $m_{00}$  give information about the degree of similarity between the two vectors  $p$  and  $q$ , whereas the entries  $m_{10}$  and  $m_{01}$  give information about their dissimilarity.

We will consider the following collections of similarity measures.

### 5.1. Similarity measures based on the Phi-similarity correlation coefficient

We start our analysis with Phi-similarity correlation coefficient (or Yule correlation coefficient, or Pearson correlation coefficient) which is given by

$$\tilde{S}_{Phi} = \frac{m_{11} \cdot m_{00} - m_{10} \cdot m_{01}}{\sqrt{(m_{11} + m_{10})(m_{01} + m_{00})(m_{11} + m_{01})(m_{10} + m_{00})}} = \frac{m_{11} \cdot m_{00} - m_{10} \cdot m_{01}}{\sqrt{p_1 \cdot p_0 \cdot q_1 \cdot q_0}}.$$

Note that the Phi-similarity coefficient is not defined when  $p_1 = 0$ ,  $p_0 = 0$ ,  $q_1 = 0$  or  $q_0 = 0$ .

The correlation measure  $\tilde{S}_{Phi}$  is sometimes attributed to Yule (1912) and it has values between -1 and 1, but it is equivalent to the formula that is obtained when the Pearson correlation coefficient (1913) originally derived for continuous data is applied to binary data.

Yule also proposed two correlation coefficients:

(1) The first correlation coefficient

$$\tilde{S}_{Yule2} = \frac{m_{11} \cdot m_{00} - m_{10} \cdot m_{01}}{m_{11} \cdot m_{00} + m_{10} \cdot m_{01}} = \frac{\frac{m_{11} \cdot m_{00}}{m_{10} \cdot m_{01}} - 1}{\frac{m_{11} \cdot m_{00}}{m_{10} \cdot m_{01}} + 1} \quad (1900).$$

(2) The second correlation coefficient

$$\tilde{S}_{Yule2} = \frac{\sqrt{m_{11} \cdot m_{00}} - \sqrt{m_{10} \cdot m_{01}}}{\sqrt{m_{11} \cdot m_{00}} + \sqrt{m_{10} \cdot m_{01}}} = \frac{\frac{\sqrt{m_{11} \cdot m_{00}}}{\sqrt{m_{10} \cdot m_{01}}} - 1}{\frac{\sqrt{m_{11} \cdot m_{00}}}{\sqrt{m_{10} \cdot m_{01}}} + 1} \quad (1912).$$

Of course, we have the following similarity coefficients as a product of the Yule correlation coefficients:

$$(1) S_{Y1} = \frac{|m_{11} \cdot m_{00} - m_{10} \cdot m_{01}|}{\sqrt{p_1 \cdot p_0 \cdot q_1 \cdot q_0}}.$$

$$(2) S_{Y2} = \frac{|m_{11} \cdot m_{00} - m_{10} \cdot m_{01}|}{m_{11} \cdot m_{00} + m_{10} \cdot m_{01}}.$$

$$(3) S_{Y3} = \frac{|\sqrt{m_{11} \cdot m_{00}} - \sqrt{m_{10} \cdot m_{01}}|}{\sqrt{m_{11} \cdot m_{00}} + \sqrt{m_{10} \cdot m_{01}}}.$$

Let us consider the Phi-similarity correlation coefficient  $\tilde{S}_{Phi}$  and remove  $m_{10} \cdot m_{01}$  from the numerator. As a result we generate a similarity coefficient

$$S_{SS} = \frac{m_{11} \cdot m_{00}}{\sqrt{(m_{11} + m_{10})(m_{01} + m_{00})(m_{11} + m_{01})(m_{10} + m_{00})}} = \frac{m_{11} \cdot m_{00}}{\sqrt{p_1 \cdot p_0 \cdot q_1 \cdot q_0}}.$$

This is a new similarity coefficient and it is called a Sokal-and-Sneath similarity coefficient (1963).

We also use the following similarly coefficients:

(1) The Peirce-1 correlation coefficient (1884)

$$\tilde{S}_{P1} = \frac{m_{11} \cdot m_{00} - m_{10} \cdot m_{01}}{p_1 \cdot q_1}.$$

(2) The Peirce-2 correlation coefficient (1884)

$$\tilde{S}_{P2} = \frac{m_{11} \cdot m_{00} - m_{10} \cdot m_{01}}{p_0 \cdot q_0}.$$

(3) The Doolittle similarity coefficient (1885) (or Pearson similarity coefficient (1926))  $S_{Doo} = \frac{(m_{11} \cdot m_{00} - m_{10} \cdot m_{01})^2}{p_1 \cdot p_0 \cdot q_1 \cdot q_0}.$

(4) The Cole-1 correlation coefficient (1949)

$$\tilde{S}_{C1} = \frac{m_{11} \cdot m_{00} - m_{10} \cdot m_{01}}{p_1 \cdot q_0}.$$

(5) The Cole-2 correlation coefficient (1949)

$$\tilde{S}_{C2} = \frac{m_{11} \cdot m_{00} - m_{10} \cdot m_{01}}{p_0 \cdot q_1}.$$

(6) The Cohen correlation coefficient (1960) (or Maxwell-and-Pilliner correlation coefficient (1968))  $\tilde{S}_{Cohen} = \frac{2(m_{11} \cdot m_{00} - m_{10} \cdot m_{01})}{p_1 \cdot q_0 + p_0 \cdot q_1}.$

## 5.2. Similarity measures based on the simple matching coefficient

The well-known Hamming distance measure for binary data  $p$  and  $q$  can be written as  $d(p, q) = m_{10} + m_{01}$ , which is also the absolute distance between  $p$  and  $q$  because we have  $\sum_{k=1}^m |p_k - q_k| = m_{10} + m_{01}$ . Obviously, the distance  $d(p, q) = \sqrt{m_{10} + m_{01}}$  is also a Hamming distance measure corresponding to the Euclidean distance. Now, using the Hamming distance  $d(p, q) = m_{10} + m_{01}$  we get the simple matching coefficient (or the simplest similarity coefficient, or M-coefficient, affinity index, or Sokal-and-Michener measure coefficient (1958), or Rand measure coefficient (1971)) which is given by

$$S_{SMC} = 1 - \frac{d(p, q)}{m} = \frac{m_{11} + m_{00}}{m} = \frac{m_{11} + m_{00}}{m_{11} + m_{10} + m_{01} + m_{00}}.$$

The corresponding Hamming dissimilarity measure is  $D_{Ham} = \frac{d(p,q)}{m} = \frac{m_{10} + m_{01}}{m_{11} + m_{10} + m_{01} + m_{00}}$ . Of course, we always have that  $S_{SMC} + D_{Ham} = 1$ .

As a consequence of  $S_{SMC}$  we obtain one new binary measure, it is also called the Russell-and-Rao measure coefficient (1940), as follows

$$S_{RR} = \frac{m_{11}}{m_{11} + m_{10} + m_{01} + m_{00}} = \frac{m_{11}}{m}, \quad \text{and its equivalent}$$

$$S'_{RR} = \frac{m_{00}}{m_{11} + m_{10} + m_{01} + m_{00}} = \frac{m_{00}}{m}. \quad \text{It is easy to show that}$$

$$S_{SMC} = S_{RR} + S'_{RR}.$$

Let us fix two positive arbitrary numbers  $a$  and  $b$ . We can define a collection of similarity measures  $S(a, b) = \frac{a(m_{11} + m_{00})}{a(m_{11} + m_{00}) + b(m_{10} + m_{01})}$  when one chooses the two real parameters  $a, b \in R_+$ . For example, when  $a=1$  and  $b=2$  we get the Rogers-and-Tanimoto similarity coefficient  $S_{RT} = \frac{m_{11} + m_{00}}{m_{11} + m_{00} + 2(m_{10} + m_{01})}$ .

### 5.3. Similarity measures based on the Jaccard similarity coefficient

Another commonly used measure is the Jaccard similarity coefficient  $S_{Jac} = \frac{m_{11}}{m_{11} + m_{10} + m_{01}}$  (1912) when  $m_{00} \neq m$ , and its

equivalent  $S'_{Jac} = \frac{m_{00}}{m_{10} + m_{01} + m_{00}}$ . As a result we also get the Hawkins-and-Dotson similarity coefficient

$$S_{HD} = \frac{1}{2}(S_{Jac} + S'_{Jac}) = \frac{1}{2} \left( \frac{m_{11}}{m_{11} + m_{10} + m_{01}} + \frac{m_{00}}{m_{10} + m_{01} + m_{00}} \right) \quad (1968) \text{ as}$$

the arithmetic mean of  $S_{Jac} = \frac{m_{11}}{m_{11} + m_{10} + m_{01}}$  and

$$S'_{Jac} = \frac{m_{00}}{m_{10} + m_{01} + m_{00}}.$$

Note that the Jaccard similarity coefficient  $S_{Jac}$  does not include the quantity  $m_{00}$ . If we consider the simple matching coefficient  $S_{SMC} = \frac{m_{11} + m_{00}}{m_{11} + m_{10} + m_{01} + m_{00}}$  and remove  $m_{00}$ , then we obtain the Jaccard similarity coefficient  $S_{Jac}$ . The Jaccard similarity coefficient  $S_{Jac} = \frac{m_{11}}{m_{11} + m_{10} + m_{01}}$  is called asymmetric because we assume that 1 is more important than 0, but we assume that 0 is more important than 1 in the similarity coefficient  $S'_{Jac} = \frac{m_{00}}{m_{10} + m_{01} + m_{00}}$ .

Another form of the Jaccard similarity coefficient is  $S_{Jac} = \frac{m_{11}}{m - m_{00}}$ .

Here the quantities  $m_{11}$  and  $m_{00}$  are very important. If  $m = m_{00}$ , then  $m - m_{00} = 0$  and  $m_{11} = 0$ , and the Jaccard similarity coefficient is not defined or indeterminate.

The Jaccard dissimilarity coefficient is given as  $D_{Jac} = 1 - S_{Jac} = \frac{m_{10} + m_{01}}{m_{11} + m_{10} + m_{01}}$ .

We also use the following similarity coefficients:

(1) The Dice (1945) (or Czekanowski, or Sorenson (1948), or Nei-and-Li (1979)) similarity coefficient  $S_{Dice} = \frac{2m_{11}}{2m_{11} + m_{10} + m_{01}}$ .

(2) The Jaccard-3 similarity coefficient  $S_{Jac3} = \frac{3m_{11}}{3m_{11} + m_{10} + m_{01}}$ .

(3) The Sokal-and-Sneath similarity coefficient  $S_{SS} = \frac{m_{11}}{m_{11} + 2(m_{10} + m_{01})}$ .

Let us fix two positive arbitrary numbers  $a$  and  $b$ . We can define a collection of similarity measures  $S(a, b) = \frac{a.m_{11}}{a.m_{11} + b(m_{10} + m_{01})}$  when one chooses the two real parameters  $a, b \in R_+$ .

#### 5.4. Similarity measures based on parts of the contingency table

Let us consider the fractions  $\frac{m_{11}}{p_1}$ ,  $\frac{m_{11}}{q_1}$ ,  $\frac{m_{00}}{p_0}$  and  $\frac{m_{00}}{q_0}$ .

Using these fractions we generate the following similarity coefficients:

(1) The Dice (1945) similarity coefficients  $S_{Dice1} = \frac{m_{11}}{p_1}$  and

$$S_{Dice2} = \frac{m_{11}}{q_1}.$$

(2) The Kulczynski (1927) (or Driver-and-Kroeber (1932)) similarity coefficient  $S_{Kul} = \frac{1}{2} \left( \frac{m_{11}}{p_1} + \frac{m_{11}}{q_1} \right)$  as the arithmetic mean of  $\frac{m_{11}}{p_1}$  and  $\frac{m_{11}}{q_1}$ .

(3) The Driver-and-Kroeber (1932) (or Ochiai (1957), or Fowlkes-and-Mallows (1983)) similarity coefficient  $S_{DK} = \sqrt{\frac{m_{11}}{p_1} \cdot \frac{m_{11}}{q_1}} = \frac{m_{11}}{\sqrt{p_1 \cdot q_1}}$  as

the geometric mean of  $\frac{m_{11}}{p_1}$  and  $\frac{m_{11}}{q_1}$ .

(4) The Dice (1945) (or Gleason (1920)) similarity coefficient  $S_{Dice} = \frac{2}{\left( \frac{m_{11}}{p_1} \right)^{-1} + \left( \frac{m_{11}}{q_1} \right)^{-1}} = \frac{2}{\frac{p_1}{m_{11}} + \frac{q_1}{m_{11}}} = \frac{2m_{11}}{p_1 + q_1} = \frac{2m_{11}}{2m_{11} + m_{10} + m_{01}}$  as

the harmonic mean of  $\frac{m_{11}}{p_1}$  and  $\frac{m_{11}}{q_1}$ . Historically, this special method for generating of the Dice similarity coefficient had been proposed by Gleason in 1920.

(5) The Braun-Blanquette (1932) similarity coefficient

$$S_{BBl} = \min \left( \frac{m_{11}}{p_1}, \frac{m_{11}}{q_1} \right) = \frac{m_{11}}{\max(p_1, q_1)}.$$

(6) The Simpson (1943) similarity coefficient

$$S_{Sim} = \max \left( \frac{m_{11}}{p_1}, \frac{m_{11}}{q_1} \right) = \frac{m_{11}}{\min(p_1, q_1)}.$$



(7) The Sokal-and-Sneath-3 (1963) similarity coefficient  $S_{SS3} = \frac{1}{4} \left( \frac{m_{11}}{p_1} + \frac{m_{11}}{q_1} + \frac{m_{00}}{p_0} + \frac{m_{00}}{q_0} \right)$  as the arithmetic mean of  $\frac{m_{11}}{p_1}$ ,  $\frac{m_{11}}{q_1}$ ,  $\frac{m_{00}}{p_0}$  and  $\frac{m_{00}}{q_0}$ .

(8) The Sokal-and-Sneath-4 (1963) similarity coefficient  $S_{SS4} = \sqrt[4]{\frac{m_{11}}{p_1} \cdot \frac{m_{11}}{q_1} \cdot \frac{m_{00}}{p_0} \cdot \frac{m_{00}}{q_0}} = \frac{\sqrt{m_{11} \cdot m_{00}}}{\sqrt[4]{p_1 \cdot q_1 \cdot p_0 \cdot q_0}}$  as the geometric mean of  $\frac{m_{11}}{p_1}$ ,  $\frac{m_{11}}{q_1}$ ,  $\frac{m_{00}}{p_0}$  and  $\frac{m_{00}}{q_0}$ . As a result we also get the Sokal-and-Sneath-

5 similarity coefficient  $S_{SS5} = S_{SS4}^2 = \frac{m_{11} \cdot m_{00}}{\sqrt{p_1 \cdot q_1 \cdot p_0 \cdot q_0}}$ .

## 6. Uncorrelatedness and statistical independence

It is well-known that statistical independence implies uncorrelatedness, but in general the converse does not hold.

In statistical theory two binary variables  $X$  and  $Y$  are uncorrelated if and only if they share zero covariance.

Now, let us study the following correlation coefficients

$$\tilde{S}_{Phi} = \frac{m_{11} \cdot m_{00} - m_{10} \cdot m_{01}}{\sqrt{p_1 \cdot p_0 \cdot q_1 \cdot q_0}}, \quad \tilde{S}_{Yule2} = \frac{m_{11} \cdot m_{00} - m_{10} \cdot m_{01}}{m_{11} \cdot m_{00} + m_{10} \cdot m_{01}} \quad \text{and}$$

$$\tilde{S}_{Yule2} = \frac{\sqrt{m_{11} \cdot m_{00}} - \sqrt{m_{10} \cdot m_{01}}}{\sqrt{m_{11} \cdot m_{00}} + \sqrt{m_{10} \cdot m_{01}}}. \quad \text{Obviously, the following statements are}$$

equivalent:

(1) The binary variables  $X$  and  $Y$  are uncorrelated.

$$(2) \tilde{S}_{Phi} = \frac{m_{11} \cdot m_{00} - m_{10} \cdot m_{01}}{\sqrt{p_1 \cdot p_0 \cdot q_1 \cdot q_0}} = 0.$$

$$(3) \tilde{S}_{Yule2} = \frac{m_{11} \cdot m_{00} - m_{10} \cdot m_{01}}{m_{11} \cdot m_{00} + m_{10} \cdot m_{01}} = 0.$$

$$(4) \tilde{S}_{Yule2} = \frac{\sqrt{m_{11} \cdot m_{00}} - \sqrt{m_{10} \cdot m_{01}}}{\sqrt{m_{11} \cdot m_{00}} + \sqrt{m_{10} \cdot m_{01}}} = 0$$

$$(5) m_{11} \cdot m_{00} - m_{10} \cdot m_{01} = 0.$$

Recall that two binary variables  $X$  and  $Y$  are statistically independent if and only if  $c^2 = \frac{m(m_{11} \cdot m_{00} - m_{10} \cdot m_{01})}{\sqrt{p_1 \cdot p_0 \cdot q_1 \cdot q_0}} = 0$ , i.e.

$$m_{11} \cdot m_{00} - m_{10} \cdot m_{01} = 0.$$

Note that for every two binary variables  $X$  and  $Y$   $c^2$ -statistics is related to the Phi-similarity coefficient by

$$\tilde{S}_{Phi} = \sqrt{\frac{c^2}{m}} = \frac{m_{11} \cdot m_{00} - m_{10} \cdot m_{01}}{\sqrt{p_1 \cdot p_0 \cdot q_1 \cdot q_0}}.$$

Finally, we deduce that uncorrelatedness and statistical independence are equivalent for every two binary variables.

## 7. Some characteristics of the similarity measures

Let us analyze the simple matching coefficient

$$S_{SMC} = \frac{m_{11} + m_{00}}{m_{11} + m_{10} + m_{01} + m_{00}} \text{ and the Jaccard similarity coefficient}$$

$$S_{Jac} = \frac{m_{11}}{m_{11} + m_{10} + m_{01}}. \text{ It is easy to prove that the following statements}$$

are true:

$$(1) S_{SMC} = \frac{m_{11} + m_{00}}{m_{11} + m_{10} + m_{01} + m_{00}} \geq \frac{m_{11}}{m_{11} + m_{10} + m_{01}} = S_{Jac}.$$

(2) If  $m_{00} = 0$ , then

$$S_{SMC} = \frac{m_{11} + m_{00}}{m_{11} + m_{10} + m_{01} + m_{00}} = \frac{m_{11}}{m_{11} + m_{10} + m_{01}} = S_{Jac}.$$

(3) If  $m_{00} > 0$  and  $m_{10} + m_{01} = 0$ , then

$$S_{SMC} = \frac{m_{11} + m_{00}}{m_{11} + m_{10} + m_{01} + m_{00}} = 1 = \frac{m_{11}}{m_{11} + m_{10} + m_{01}} = S_{Jac}.$$

(4) If  $m_{00} > 0$  and  $m_{10} + m_{01} > 0$ , then

$$S_{SMC} = \frac{m_{11} + m_{00}}{m_{11} + m_{10} + m_{01} + m_{00}} > \frac{m_{11}}{m_{11} + m_{10} + m_{01}} = S_{Jac}.$$

(5) If  $S_{SMC} = \frac{m_{11} + m_{00}}{m_{11} + m_{10} + m_{01} + m_{00}} = \frac{m_{11}}{m_{11} + m_{10} + m_{01}} = S_{Jac}$ , then

$$m_{00} \cdot (m_{10} + m_{01}) = 0, \text{ i.e. } m_{00} = 0 \text{ or } m_{10} + m_{01} = 0. \text{ See also (2) and (3).}$$

## References

- [1] Cha S., S. Yoon, C. Tappert, Enhancing Binary Feature Vector Similarity Measure using Genetic Algorithm, *Journal of Pattern Recognition Research*, 2006.
- [2] Choi S., S. Cha, C. Tappert, A Survey of Binary Similarity and Distance Measure, *Journal of Systemic, Cybernetics and Informatics* (2010) 8, 43-48.
- [3] Goodman L., W. Kruskal, Measures association for cross classifications, *Journal of American Statistical Association* (1954) 49, 732-764.
- [4] Goodman L., W. Kruskal, Measures association for cross classifications II. Further discussion and references, *Journal of American Statistical Association* (1959) 54, 123-163.
- [5] Goodman L., W. Kruskal, Measures association for cross classifications III. Approximate sampling theory, *Journal of American Statistical Association* (1963) 58, 310-364.
- [6] Cormack R., A review of classification, *Journal of the Royal Statistical Society – Series A* (1971) 134, 321-353.
- [7] Forbes S., Method of determining and measure the associative relations of species, *Science* (1925) 61, 524-537.
- [8] Hubalek Z., Coefficients of Association and Similarity, Based on Binary (Presence-Absence) Data: An Evaluation, *Biological Reviews* (1982) 57-4, 669-689.
- [9] Jackson D., K. Somers, H. Harvey, Similarity Coefficients: Measure of Co-Occurrence and Association or Simply Measures of Occurrence, *American Naturalist* (1989) 133-3, 436-453.
- [10] Michael E., Marine ecology and the coefficient of association: a plea in behalf of quantitative biology, *Ecology* (1920) 8, 54-59.
- [11] Michael H., Binary coefficients: A theoretical and empirical study, *Mathematical Biology* (1976) 8-2.
- [12] Slavov Z., Mathematical Ideas in Centroids-Based Partitioning Cluster Analysis, *International Conference ISK 2008, Varna 26 – 28 June 2008*, 336-345.
- [13] Slavov Z., On the Centroids-Based Partitioning Cluster Analysis in Mechanical Industry, *International Conference RaDMI 2008, Uzice 14-17 September 2008, Serbia*, 1-7.
- [14] Slavov Z., A Survey of Similarity Measure in Statistical Analysis of Binary Data, *ISC "Statistics as a science, education, profession and activity, Varna 24 - 25 October 2013*, 57-61.
- [15] Sneath P., R. Sokal, *Numerical Taxonomy: Principles and Practice of Numeric Taxonomy*, San Francisco, W.H.Freeman and Company, 1973.
- [16] Sokal R., P. Sneath, *Principles of numeric taxonomy*, San Francisco, W.H.Freeman and Company, 1963.

[17] Willett P., J. Barnard, G., Downs, Chemical similarity searching, Chem Inf Comput Sci (1998) 38, 983-996.

[18] Willett P., Similarity- based approaches to virtual screening, Biochemical Society Transactions (2003) 31, 603-606.

**Authors:**

Prof. **Zdravko Dimitrov Slavov**, PhD

Varna Free University, Bulgaria

[slavovibz@yahoo.com](mailto:slavovibz@yahoo.com)

**Desislava Zdravkova Slavova**

Virginia Tech, Blacksburg, USA

[slavova\\_desi@yahoo.com](mailto:slavova_desi@yahoo.com)

**INFORMATICS IN THE SCIENTIFIC KNOWLEDGE**

**International Scientific Conference**

Managing editor *Prof. Dr. Eng. Nikolay Lyutov*

Cover designer *Nikolay Dimitrov*

Bulgarian, 1st edition

Format 70/100/16

ISSN 1313–4345

© University publishing house  
VFU “Chernorizets Hrabar”