International Conference

Big Data, Knowledge and Control Systems Engineering

BdKCSE'2016

Sofia, Bulgaria
1-2 December 2016

Institute of Information and Communication Technologies
- Bulgarian Academy of Sciences

John Atanasoff Society of Automatics and Infomatics

PROCEEDINGS

International Conference on

Big Data, Knowledge and Control Systems Engineering - BdKCSE'2016

1-2 December 2016
108 G. S. Rakovski Str., Hall 105A, 1000 Sofia, Bulgaria

Institute of Information and Communication Technologies
of the Bulgarian Academy of Sciences
"John Atanasoff" Union on Automatics and Informatics, Bulgaria

Editor:
Rumen D. Andreev

Cover Designer: Ivan Panayotov
Prepress: Ivan Panayotov

Department of Communication Systems and Services
Institute of Information and Communication Technologies - Bulgarian Academy of Sciences
Bl. 2, Acad. G. Bonchev Str., 1113 Sofia, Bulgaria

Conference scope

The International Conference "Big Data, Knowledge and Control Systems Engineering" (BdKCSE'2016) aims to provide an open forum for the dissemination of the current research progress, innovative approaches and original research results on all aspects of Big Data Management, Technologies, and Applications. Organizer of the BdKCSE'2016 Conference is the Institute of Information and Communication Technologies of the Bulgarian Academy of Sciences, and co-organizer is the "John Atanasoff" Union of Automatics and Informatics, Bulgaria.

Big Data Management, Technologies, and Applications discuss the exponential growth of information size and the innovative methods for data capture, storage, sharing, and analysis. Modern technologies continue to become more complex as do the applications. The integration of technologies, complex relationships of applications and the accelerated technological changes are new challenges to technology management.

Topics such as product development, innovation management, and research and development management have become very popular. Big data spans dimensions as volume, variety, velocity, volatility and veracity, steered towards one critical destination – value. Following from these, the conference is devoted toward improving the understanding, systems engineering, human cognition and modeling, and data.

The conference will help the research community identify the novel important contributions and opportunities for recent research on the different intelligent methodologies and techniques in the field.

Table of contents

Organized by:

Institute of Information and Communication Technologies
- Bulgarian Academy of Sciences

"John Atanasoff" Union on Automatics and Informatics, Bulgaria

Our conference will take place at the
Federation of the Scientific Engineering Unions /**FNTS**/, Bulgaria

Program committee

Conference Chairs

- Chairman Assoc. Prof. Rumen Andreev Bulgarian Academy of Sciences
- Vice chairman Assoc. Prof. Lyubka Doukovska Bulgarian Academy of Sciences
- Vice chairman Assoc. Prof. Yuri Pavlov Bulgarian Academy of Sciences

Program Committee

- Abdel-Badeeh Salem Ain Sham University Egypt
- Chen Song Xi Iowa State University USA
- Dimiter Velev University of National and World Economy Bulgaria
- František Čapkovič Slovak Academy of Sciences Slovakia
- George Boustras European University Cyprus
- Georgi Mengov University of Sofia Bulgaria
- Jacques Richalet France
- John Wang Montclair State University USA
- Kosta Boshnakov University of Chemical Technology and Metallurgy Bulgaria
- Krasen Stanchev University of Sofia Bulgaria
- Ljubomir Jacić Technical College Požarevac Serbia
- Ljupco Kocarev Macedonian Academy of Sciences and Arts Macedonia
- Milan Zorman University of Maribor Slovenia
- Neeli R. Prasad Aalborg University, Princeton USA
- Olexandr Kuzemin Kharkov National University of Radio Electronics Ukraine
  German Academic Exchange Service Bonn
  North Rhine-Westphalia Germany
- Pedja Milosavljević University of Niš Serbia
- Peter Kokol University of Maribor Slovenia
- Petko Ruskov University of Sofia
- Radoslav Pavlov IMI, Bulgarian Academy of Sciences Bulgaria
- Rumen Nikolov UniBIT-Sofia Bulgaria
- Song II-Yeol Drexel University USA
- Sotir Sotirov University “Prof. Asen Zlatarov” Bulgaria
- Tomoko Saiki Tokyo Institute of Technology Japan
- Uğur Avdan Anadolu University Turkey
- Valentina Terzieva IICT, Bulgarian Academy of Sciences Bulgaria
- Valeriy Perminov National Research Tomsk Polytechnic University Russia
- Vera Angelova IICT, Bulgarian Academy of Sciences Bulgaria
- Vyacheslav Lyashenko Kharkov National University of Radio Electronics Ukraine
- Wojciech Piotrowicz University of Oxford UK
- Zlatogor Minchev IICT, Bulgarian Academy of Sciences Bulgaria
- Zlatolilia Ilcheva IICT, Bulgarian Academy of Sciences Bulgaria

Publisher:
"John Atanasoff" Union on Automatics and Informatics, Bulgaria

108 G. S. Rakovski Str., 1000 Sofia, Bulgaria
Phone: +3592 987 6169
e-mail: sai.bg.office@gmail.com
www.sai.bg

ISSN: 2367 - 6450

Industrial Process Optimization through Quality Tools

Pedja Milosavljević, Milena Rajić, Dragan Pavlović, Ana Petrović
University of Niš, Faculty of Mechanical Engineering, Department of Management in Mechanical Engineering
Aleksandra Medvedeva 14, 18000 Niš, Republic of Serbia
pedja@masfak.ni.ac.rs, milenatod1@yahoo.com, draganpavlovic10369@gmail.com

Abstract: The aim of this paper is to identify all of the defects that occur as losses and complicate the process of production in order to achieve the Lean Six Sigma level in the Shinwon Company. The production process in the company must constantly be analyzed, improved and sustained. This paper outlines the losses in production process of the company, but also necessary measurements that need to be applied in order to improve the processes of the company. The original data from the Shinwon company were identified, collected and analyzed in order to present the effectiveness of the quality management system and to evaluate the possibility of its continuous improvement. Following the acquisition, the data were analyzed using the methods and tools of the Lean Six Sigma concept (5S audit, Pareto diagram, Ishikawa diagram, Seven basic wastes), in order to improve the quality of business operations. After analyzing the data, certain improvements were proposed in order to elevate the entire enterprise to the Lean Six Sigma level of organization.

Keywords: Lean Six Sigma Method, Industrial process improvement, Quality tools, Pareto diagram.

1 Introduction

The rapid development of the market has imposed the obligation to companies to introduce permanent improvements in their systems and to train their employees, in order to become competent to introduce the new methods in the area of process quality improvement.

Improvement of the processes has become an important factor in gaining an advantage over competitors. In the higher profit race and struggling for survival in times of global crisis, there is less money and available time for improvement, so therefore new ideas are needed to be introduced.

Response to the new circumstances on the market can be found in Lean Six Sigma concept. Lean Six Sigma concept demands constant changes and constant improvement. The emphasis is on employees involvement and teamwork, measurement and systematization of processes, reducing variation, defects and shortening the duration of the process.

Therefore Lean Six Sigma is a combination of the two most important trends to develop and improve the working results and make the work better (through Six Sigma) and make it faster (through Lean principles).

Production process in company Shinwon is also based on Lean Six Sigma concept. The production process in the company should be constantly analyzed and improved. This paper outlines all the shortcomings, i.e. losses that hamper the production process of this company.

2 Seven Basic Types of Waste

Waste is everything that adds cost or is time-consuming without adding the value to the process [3], [6]. Each activity in the company spends some resources that can always be converted into the money. The goal of Lean enterprise is to eliminate any unnecessary activities that do not contribute to the product value.

There are 7 basic types of waste, defined by the Toyota managers (over-production, transport, movements, waiting, over-processing, inventory, defects), which can also be applied in the company Shinwon. Such systematization can be applied in any company, to any process and it is a basic of Lean concept - company without losses. Based on the production process in the company Shinwon it can be seen which types of the waste exist.

Table 1 - Technical characteristics of the boilers

No.	**Waste**	**Definition**
1	Defects	This type of waste is the most common in the company Shinwon and represents the biggest problem that occurs during the production process. Defects can occur during production or they can be detected when the products come to the customer.
2	Inventory	This type of waste does not occur, because the company does not have unnecessary inventories.
3	Movements	There is no waste in this category, each of the workers are in their working positions and the leader of production takes the material from the storage of materials needed for the production process.
4	Waiting	Waiting is possible for certain semi-material (set of wires) supplied by the company Yura Corporation from Niš.

5	Transport	There is no waste for this category. Transport of materials, semi-finished and finished products in the company Shinwon is minimized.
6	Over-processing	This type of waste is very common in the company Shinwon. Every day, the sector of quality control discovers errors, ie. defects that occur on the installation during its production. In that case over-processing is necessary, thereby increasing production time.
7	Over-production	In this category we cannot identify any waste. The entire production is planned and implemented based on pre-defined customer requirements, according to the monthly, weekly and daily orders, so there is no excess production.

3 Pareto Diagram

Pareto diagram or ABC diagram, named after the Italian economist Vilfredo Pareto, is a tool that is used to identify and group causes of problems according to their relative importance [4, 5]. This quality tool is used in cases where it is possible to identify errors, their frequency and / or created expenses and to take corrective actions in order to eliminate errors. In other words, it represents the process of selecting priority issues for solving, and it is used to focus on the vital minority (20%), which leads to significant improvement (80%) [2]. By introducing appropriate corrective and preventive measures, it's possible to prevent that problems never occurr in the future [7].

Pareto diagram for the production process of installation is done by monitoring the defects (errors) that occurred during the process. The data are given for year 2015. Defects or errors may occur during both the production process - on the SUB - the preparation of wire for the further production process, and ASSEMBLY line - during assembly of installation. Those errors can be detected by the quality control during the installation test on both control tables - electrical test control and visual inspection test. Electrical test control or test of electrical safety of installations detects defects on installation, which are results of workers mistakes on SUB, while table for visual inspection detects defects made during assembly of installations.

Pareto diagram for the production process of the installation shows wastes in the form of defects that are most common and most serious problems in the company Shinwon. This requires improvement in order to increase the effectiveness and efficiency of the process. Data

are based on the occurrence of 10 most common defects of installations that were reported on the control electric table in year 2015 (figure 1).

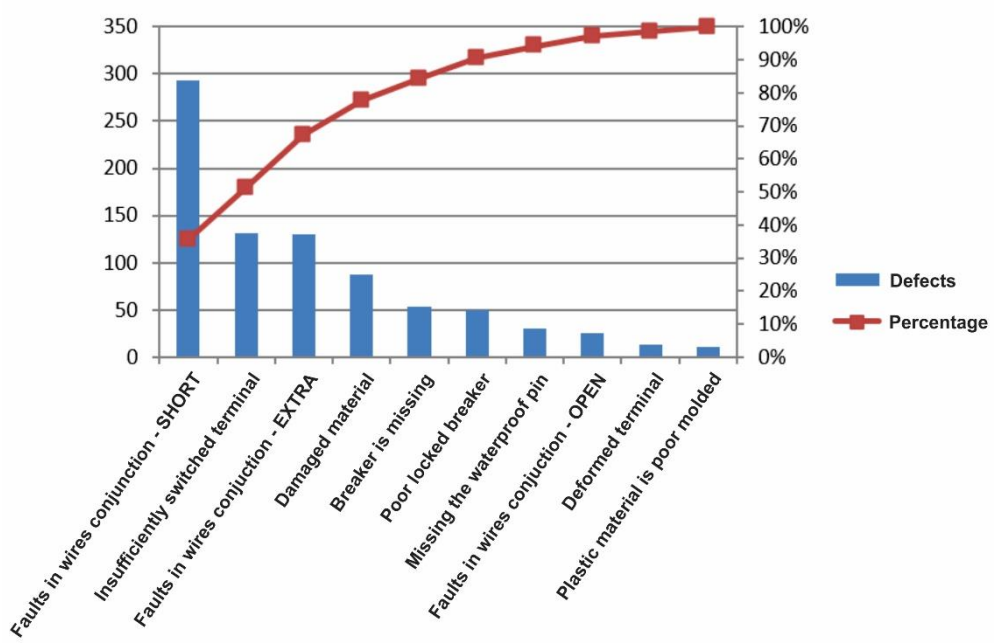


Figure 1 – Pareto diagram of installation defects on the control electric table

Pareto diagram shows that the biggest problem in the process are defects, ie. mistakes in conjunction of wires - SHORT. This error occurs due to negligence of workers on the SUB or their insufficient or inadequate education. This is an error that occurs by replacing the wires in the connector. It may happen that workers turn the connector upside down and thus leads to defective installation, or they are use incorrectly the drawings, patterns and/or work instructions.

Based on data for the 2015th year, the 10 most common defects on installations, which are discovered on the table for visual control, are shown on the Pareto diagram below (figure 2).

Pareto diagram shows that in most cases the process slows down due to an error in the dimension of installation and such a defect represents a serious problem to further correction of entire installation. This error occurs during the assembly (bandages) of installation on the ASSEMBLY line, and the defect may be in the form of short and/or long installation. Short - if the installation is too much wrapped, ie. too much insulating tape on the wire. Longer - if the wire is not enough wrapped with insulating tape. This error occurs because the workers on

the assembly line do not follow the work instructions, which are located on the assembly board. This is due to their negligence or inadequate training.

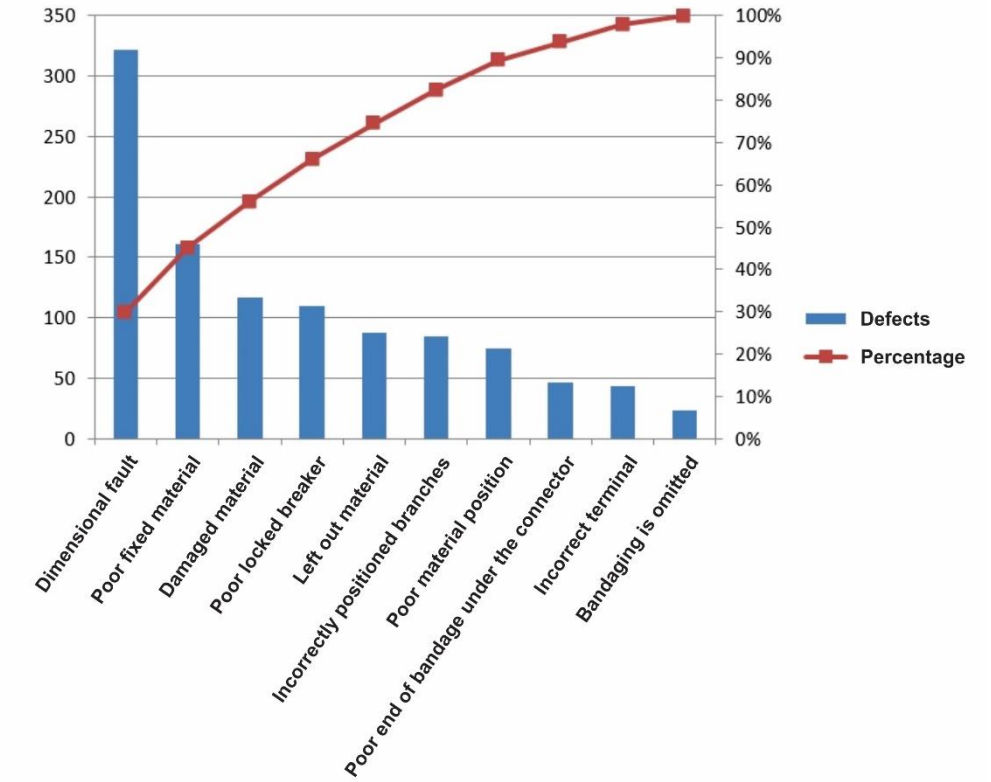


Figure 2 – Pareto diagram of installation defects on the table for visual control

4 Ishikawa Diagram

Ishikawa diagram is a tool that helps in identification, sorting, and displaying possible causes of a specific problem or quality characteristics. The diagram graphically shows the relation between specific consequence and all factors that influence the consequence [1].

Ishikawa diagram was made on the basis types of waste that occur in the production process of the company Shinwon (figures 3, 4).

The picture shows the Ishikawa diagram that systemically and structurally analyzes defects as a result which leads to maximum waste. Identified consequence is entered in the diagram on the right diagram spine. It is necessary to identify the main causes, such as machines, methods, materials and men that has effect on the consequence. These causes are

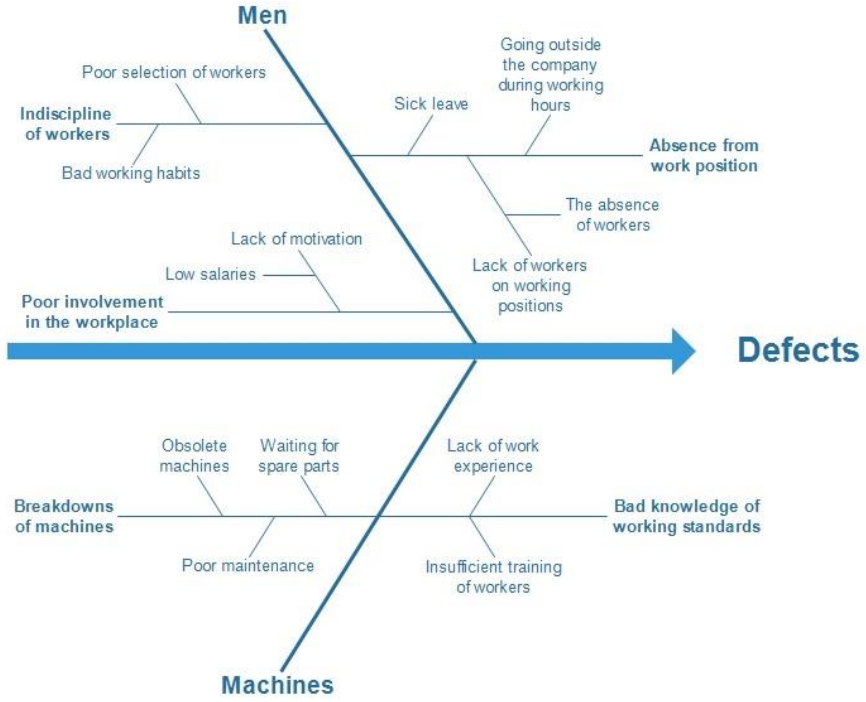


Figure 3 – Ishikawa diagram for categories Men and Machines

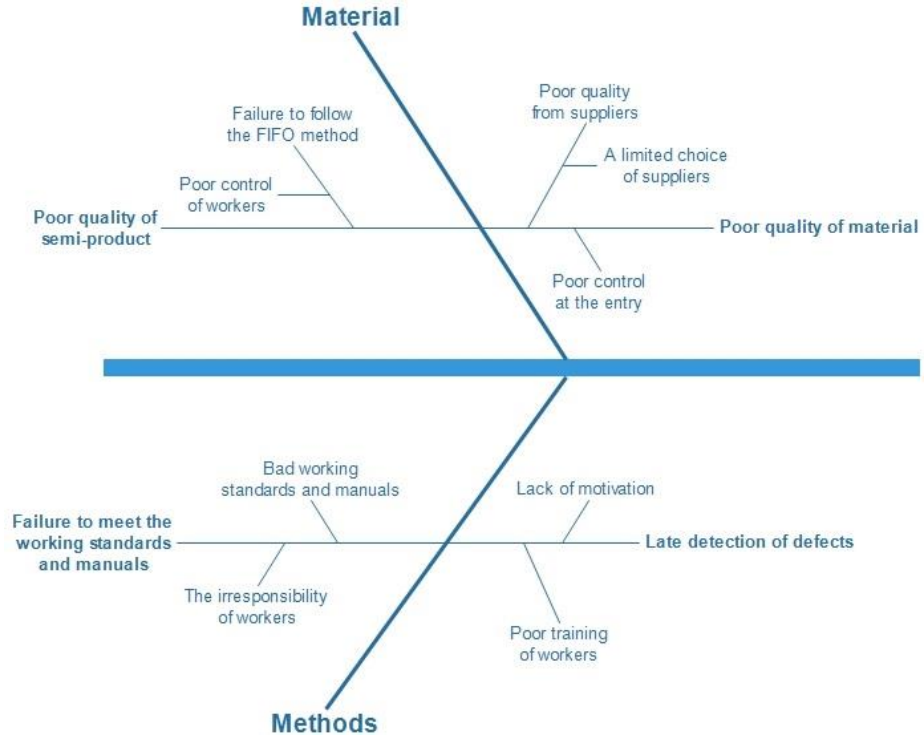


Figure 4 – Ishikawa diagram for categories Material and Methods

the main branches of the diagram. For each category, the other specific factors, which may cause consequences are identified. The analysis is used to identify the causes that justify further investigation.

5 5S Method

5S is a method of Lean organization and represent a set of rules for the workplace organization of each employee. The aim is that each workplace is organized in such way to be maximally efficient and to speed up and facilitate the work of the employee.

The method is named after the Japanese words that begin with the letter S:

- Seiri – Sort;
- Seiton – Set in order;
- Seiso – Shine;
- Seiketsu – Standardize;
- Shitsuke – Sustain.

Since its beginning Shinwon company is implementing 5S method and constantly working on its improvement. Quality control managers checks each month implementation of 5S method by using checklists. In case there is some irregularities in the implementation of 5S audit, a report of improvement (Shinwon Improvement Report) is done with a detailed description of the problem and its solution.

The following example (Figure 5) shows a 5S audit checklist for the first phase - sort and solution to all the irregularities that occur (for month May).

THE CHECKLIST FOR 5S METHOD SORT		Prepared	Checked	Apprtoved
		Date		
ACTIVITIES		**YES**		
1	Only necessary tools are present at the workplace	☑		
2	Only necessary equipment is present at the workplace	☑		
3	Only necessary materials and auxiliary equipment are present at the workplace	☑		
4	The necessary documentation is only present at the workplace	☑		
5	Secured parts and electric cables which represent a danger in production	☑		

Figure 5 – The checklist for phase Sort

The following Figure 6 shows the state of the company Shinwon during the last audit (for month May).

Figure 6 – Box covers

During the control of 5S audits some irregularities can be noticed in the production process. The Shinwon Improvement Reports is represented below and shows improvement measures for certain irregularities arise during the process of production (for month May). Irregularity which was discovered during the implementation of the first phase of the 5S method is corrected. Box covers can be damaged and scattered due to inadequate protection. Workers must use a particular location for proper storage of the box covers, which has been done.

Figure 7 – Improvement report for the first phase of 5S audit

6 Conclusion

Detailed analysis of the production process of wiring production company Shinwon is done by observing the activities that need to be improved to make the process more efficient and effective. While analyzing the list of 7 basic types of waste is concluded that in the production process comes to waiting due to delays of materials or semi-finished products (a set of wires) that delivers Yura Corporation in Nis, and that in process defects are occurring which represent the reason to process of treating or retreating the installations. These processes increasing the cost of production and may threaten the continuity of the production process. These are the three categories that are most critical to the whole process. However, the main problem, whose solution should strive, emphasizes the reporting defects during the production processes. Pareto diagrams show the 10 most common defects on installations that have been observed during 2015. The main cause of the defects is insufficient or inadequate traing of employees or their negligence during the operations.

The results show that the reduction in defects may be achieved by improving the processes in several ways: motivation for choosing the best employee of the month, payment or salaries based on employees productivity and other measures that reward may encourage employees to diligently and responsibly perform their duties; educating employees through additional training; standardization of all operations performed around the working area to be more understandable and closer to the conditions in the working environment; necessity to constantly improve every process, to hold meeting with employees in order to improve quality. Encourage employees to suggest improvements in order to implement Kaizen – continuous improvement. Every idea should be encouraged and respected.

Implementation of Lean Six Sigma concept requires time, effort and costs. Users get better product and services and become more loyal to the company and the brand, the company is profitable and offers a safer job. The implementation of Lean Six Sigma companies greatly increase efficiency and productivity, which contributes to their competitiveness in the international market.

References

[1] Ishikawa K. (1982) Guide to Quality Control, *Asian Productivity Organization*, Tokyo.

[2] Hung H. C., Sung M. H. (2011) Applying Six Sigma to Manufacturing Processes in the Food Industry to Reduce Quality Cost. *Scientific Research and Essays*, 6, 3, 580-591.

[3] Mladenovic M., Milosavljevic P. (2010) The road towards a Lean Six Sigma company. *International Journal Total Quality Management & Excellence*, 38, 3.

[4] Seferovic E., Cengic D. (1998) Alati Za Upravljanje Kvalitetom u Primjeni. *Masinstvo*, 4, 2, 193-204.

[5] Stoiljkovic V., Jankovic Z., Markovic I. (2010) Application of Pareto Analysis in Pharmaceutical Institutions. *International Journal Total Quality Management & Excellence*, 38, 3, 93-97.

[6] Stoiljkovic V., Stoiljkovic P., Stoiljkovic B. (2009) Implementation Lean Six Sigma concept in manufacturing and service organization. *International Journal Total Quality Management & Excellence*, 37, 1-2.

[7] Stoiljkovic V., Trajkovic J., Stoiljkovic B. (2010) Lean Six Sigma sample analysis process in a microbiology laboratory. *Journal of Medical Biochemistry*, 30, 4, 346

On Entropy and Concentration Analysis of Industrial and Market Systems

Iliyan Petrov (PhD in Economics)
Gubkin University for Oil and Gas,
Leninskyi pr. 65, Moscow, Russian Federation
petrovindex@gmail.com

Abstract: In this paper, after brief recapitulation classical "information entropy" and other existing tools for measuring diversity, we presents improved methods for studying concentration and hierarchy in complex systems. With original concepts for "Phased-Structure States" and "Set Concentration Index" we integrate several new indicators in a logical and reliable system for quantitative and qualitative assessment of industrial and market structures. The practical application of our innovative approach in the controversial world energy provides convincing results, which prove, that such flexible methodology is suitable for large scope of tasks of structural and risk analysis in different sectors (industrial, financial, transport, informatics, etc.) and other areas (sociology, biology, health care, ecology, etc.).

Keywords: information theory, entropy, phase-structure states, set concentration index, industry organization, market concentration, competition, cartels, monopoly power, antitrust regulation.

1 Introduction

The globalization of world economy have growing influence on industrial structures and market competition for resources, technology innovation and information networks. Attempts to abuse with privileged position or cartel interactions increase the concerns of consumers and regulators for preserving free competition and adequately regulate the activities of natural monopolies. Concentration aspects have key importance in several industry sectors (incl. energy, transport information technology, finance, etc.).

2 Understanding and Measuring Market Structures

From narrow point of view, any "structure" represents an arranged set of entities, while from dynamics point of view any set and sub-set of such entities are investigated with respect of their inputs, transformation process and corresponding output results in the system.

In broader term, the "system structure" is a complex of interrelated components forming the framework of core interactions concerning the access to resources, information and technology. **The number of entities** and their access to resources determine the level of hierarchy (concentration) or diversity (chaos, entropy) in the system. **The institutional framework** regulates the activities and interactions in the system. **The existing technological processes** determines the treatment and transportation material, financial and human resources in the system. **The information flows and systems** reflect the processes of collection, treatment and interpretation of data.

2.1 Overcoming Limitations in Existing Tools for Structural Analysis

Currently there are different concepts and *tradeoff functions* for measuring system concentration or diversification. All of use the format $\boldsymbol{F = \sum_{i=1}^{N} f(s_i)}$, where ***"N"*** stands for number of players, ***"s_i"*** - player's share in a system; $\boldsymbol{f(s_i)}$- a basic non-linear *tradeoff function*; "***F"*** – summing the information about individual entities in an ensemble.

The *Herfindahl-Hirschman Index is defined by a simple quadratic function* $\boldsymbol{HHI= \sum_{i=1}^{N}(s_i)^2}$, used officially to measure the concentration of markets and monopoly power /(Herfindahl, 1950), (Hirschman, 1945)/. Such simplistic *model* contains just one calculation, which, unfortunately, postulates to economic theory inadequate concepts for guaranteed steady growth and distortions in data filtering.

The *Shannon-Wiener Index* $\boldsymbol{SWI=H(q)=-\sum_{i}^{N}(q_i \log_2 q_i)}$, which measures the entropy of information, is considered as one of the "visit cards" for leading scientific findings of 20th century's in the area modern information technologies. However, its' parabolic and non-monotone profile turned out to be non-applicable for describing socio-economic systems.

In 1961 Alfred Renyi (Renyi, 1961) proposed his generalized approach on entropy $\boldsymbol{H_{\propto}(X) = \frac{1}{1-\propto}\log_2(\sum_{i=1}^{n} p_i^{\propto})}$**.** The limiting value of $\boldsymbol{H_{\propto}}$ as $\boldsymbol{\alpha \to 1}$ in fact leads to Shannon's concept of information entropy $\boldsymbol{H_1(X) = -\sum_{i}^{N}(p_i \log_2 p_i)}$.

Taking into account both the achievements and limits of main existing models, our research focuses on two main areas: a) improving methodology for quantitative measurement, qualitative assessment and classification of structures; b) practical application of innovative approaches for analyzing industry organization and concentration in key world markets (incl. energy sector in this paper). To improve the methodology of structural analysis we developed a set of new concepts, models and indicators.

2.2 Structural Evolution of "Phase-structural States (PhSS)"

For describing the system evolution we propose an original concept, called *Phase-Structural States (PhSS)*. Phase boundaries are determined clearly by the values of symmetric states, in which all entities have equal weights (1-0,5-0,33-0,25-0,2 ... etc.) and entity's belonging depends on its relative weight (share, ***"s_i"***) in the system**.** Thus, an entity with a share from 0.3 to 0.5 belongs to the 3-rd structural phase. To classify a system we take into account the number entities and the discrete distribution of their shares. The *PhSS concept* allows to define profiles of the *summing function* ***"$F=\sum_i^n \eta(si)$"*** for modelling extreme and moderate paths of *system evolution*. With a compact set of five logically interrelated functions we describe both the evolution of the individual entity and the key paths of *PhSS* (*minimum, maximum and average "working" states)* of the system as a whole.

The key point of any diversity index is its *basic concept model (BCM)*, which filters the information about the growth of individual entities. In our innovative model, called ***"Set Concentration Index" (SCI),*** we propose an original *tradeoff function* /***"$\eta(s_i)$"***/ with *logistic profile (S-curve type)* for defining the transformation of original data values for relative weights into singular corresponding values within a universal dimensionless scale. The *basic tradeoff function* ***SCIbas*** (equation 3, table 1) filters non-lineary the importance of an entity by comparing its' share ***("s")*** with the system size as a whole ***("unity", or "1")*** in addition with some important information about the structure a system, which is both self-organized and regulated. We are able to take into account key aspects of interactions (competition) in "producing internal entropy" (caused by internal factors) and "inducement of external entropy" (caused by external factors), (Prigozhin, 1990). As a result, the setting of *"reference structural thresholds (RST)"* reflects the influence of objective internal characteristics of systems (industries, markets, etc.) and the external regulation (administration). The optimizing of RST (as fixed parameters) in *SCIbas* leads to selecting ***variant "4-1000"***, i.e. ***SCIbas(4-1000)***, which contains only two "thresholds". The *lower RST* ***b_1=0,001*** (i.e. 0.01% share) reflects a minimal critical mass for viability as a self-regulating factor in different kind of socio-economic systems. The antitrust legislation in developed countries regards the 75% share of three market leaders (so called "CR3") as a transition zone to highly concentrated structures, which allows to derive a value for the *upper RST* as ***b_2 = 0,25*** (i.e. 25% share in a symmetrical market with four participants (4x25%=100%) (fig. 1).

With moderate level of *system interactions* (***n=2***) the central balancing value of ***SCISUMbas=0.5*** marks the transition from "less concentrated" to "more concentrated" structures, reflecting antimonopoly regulation in the EU, Japan, Russia and other countries (as

well as United States until 2010) for fully symmetrical market with 10 companies (each with equal share of 10%). The synthesis of *PhSS concept* and *SCI model* allows to develop improved indicators for measuring synergy effects from cooperation in coalitions. The ***PhSS-SCI model*** and the new indicators form an integrated ***System for Assessment and Classification of Structures (SACS),*** (table 1).

In addition, we introduce a new system for classification of market structures, comprising 5 main stages: *1) monopoly; 2) oligopoly; 3) polipoly 4) multipoly; 5) hyperpoly*. Such approach improves the analysis of oligopolistic markets, refines the understanding of less concentrated markets and sets objective criteria for the content of commonly used clichés, like "large number of market participants", "free markets" and "perfect competition".

Structures with *concentrated PhSS (SCISUMbas ∈ 0,5÷1)* are categorized in 3 stages: *concentrated oligopoly (partial monopoly), classical oligopoly and enlarged oligopoly.* Structures with *non-concentrated PhSS (SCISUMbas ∈ 0÷0,5)* are also categorized in 3 stages: *polipoly, multipoly and hyperpoly*. The profile of the summing function, which simulates *minimum concentration PhSS /SCISUMbas(simmin)/* in variant *SCIbas(4-1000)* (equation 6, tab. 1) is very similar in profile to the ***Harrington Desirability Function*** (Harrington, 1965). Moreover, due to its profile with several discrete horizontal smoothenings in the zones of *minimum PhSS*, our model offers much better possibilities for quantitative assessment and multi-level qualitative classification. In the traditional HDF concept with five evaluation intervals (1÷0,8÷0,63÷0,37÷0,2÷0) we propose to split the central interval (0,63÷0,5 ÷0,37) for obtaining six intervals, which are logically symmetric (1÷0,8÷0,63÷0,5÷0,37÷0,2 ÷0).

As a result, the central value of ***SCISUMbas=0.5*** sets a natural balance point for the qualitative transitions in all indicators: in ***SCISUMbas*** - from "less concentrated" to "more concentrated structures"; in ***CFI*** and ***PMI*** - from *competition* to *domination* type of interactions; in ***MTSI*** - from *"buyer's"* to *"seller's market"*.

The advantages of the *SCI model* are evident in comparison with the inflexibility of other indicators - the *Shannon-Wiener index (SWI*) *for information entropy* and *Herfindahl-Hirschman concentration Index (HHI)*. As novelty we use defined integrals as an elegant tools for measuring the volume of information contained the in the basic functions of index models.

Table 1. Generalization of "Structure Assessment and Classification System (SACS)" in PhSS-SCI Model

Phase-Structure States (PhSS) Model	Limited Development Space-Time Model in the Set Concentration Index (SCI) /Set Hierarchy Index (SHI)	Market Structure Type Index (MSTI) $= 0.5\,[IE(s_{iSupply}) / IE(s_{iDemand})]$	
Basic concept model (BCM)	$SCIbas(s_i)=\eta(s_i)=\frac{s_i}{1+(\sum_{j=1}^{m}\log_{b_j} s_i)/m)^n}$ (3)	Competitive Force Index (CFI)	
Universal Reference Model (URM)	If $m=n=2;\ b_1=0.1\%=0.001;\ b_2=25\%=0.25$ $SCIbas(4-1000)=\frac{s_i}{1+0.1875\,(\ln s_i)^2}$ (3.1)	**Individual:** $CFI(Supply)=\frac{\eta(s_{iSup\ ply})}{IE(s_{iSup\ ply})}$ (9.1)	**Coalition:** $CFIcoal(Supply)=\frac{\eta(s_{ilMaxSup})+k[\eta(\sum_i^l s_{ilSup})-\eta(s_{ilMaxSup})]}{IE(s_{iSupply})}$ (9.3)
Structure Assessment Model (SAM) Structure Minimal Concentration Model (SMCM)	$SCISUMbas=IEbas(s_i)=\sum_{i=1}^{N}\eta(s_i)$ (4) $IEbas(4\text{-}1000)(s_i)=\sum_{i=1}^{N}(\frac{s_i}{1+0.1875\,(\ln s_i)^2})$ (4.2) $SCISUMbas(symmin)=IE(symmin)$ (6) Discrete distribution of symmetric states with minimal concentration (hierarchy) of the structure	$CFI(Demand)=\frac{\eta(s_{iDem\ and})}{IE(s_{iDem\ and})}$ (9.2)	$CFIcoal(Demand)=\frac{\eta(s_{ilMaxDem})+k[\eta(\sum_i^l s_{ilDem})-\eta(s_{ilMaxDem})]}{IE(s_{iDemand})}$ (9.4)
Maximal Concentration Model	A. Simulation model (7.1) B. Mathematic model: when $b_o=0.01\%=0.0001;\ n=2$ $SCImax=\eta max(s_i)=\frac{s_i}{1+(\log_{b_o} s_i)^2}$ (7.2)	Market Power Index (MPI) **Individual:** $MPI(sup/Dem)=\frac{\eta(s_{iSup\ ply})}{IE(s_iDemand)}$ (10.1)	**Coalition:** $MPIcoal(sup/Dem)=\frac{\eta(s_{ilMaxSup})+k[\eta(\sum_i^l s_{ilSup})-\eta(s_{ilMaxSup})]}{IE(s_{iDemand})}$ (10.3)
Average Model - arithmetic mean - geometric mean	$IEarithmeam=(IEbas(symmin)+IEmax)/2$ (8.2) $IEgeomean(math)=\sqrt{IEbas(symmin).IEmax(math)}$ (8.2) $IEgeomean(simul)=\sqrt{IEbas(symmin).IEmax(bassimul)}$ 8.3)	$MPI(dem/Sup)=\frac{\eta(s_{iDem\ and})}{IE(s_iSupply)}$ (10.2)	$MPIcoal(dem/Sup)=\frac{\eta(s_{ilMaxDem})+k[\eta(\sum_i^l s_{ilDem})-\eta(s_{ilMaxDem})]}{IE(s_{iSupply})}$ (10.4)

Abbreviations: ***IE*** – *hierarchy of a system (from greek "ιεραρχία" – hierarchy)* ***Dem*** – *Demand;* ***Sup*** – *Supply;;* s_i – *Player's Market Share;* ***i*** – *Market Player;* ***N*** – *Players' Number;* s_{il} – *Coalition participant's Market Share;*
k – *Coalition Cooperativeness Coefficient;* ***l*** – *Coalition Players' Number;* ***il*** – *Coalition Participant* s_{ilMax} – *Market Share of Coalition Leader;* ***coal*** – *Coalition;*
b_j – *Reference Structure Threshold (RST);* ***m*** – *number of RSTs;*
n – *level of market interaction between market shares and structures threshold (**1≤n≤3**; **1** – week interaction; **2** – moderate interaction; **3** - strong interaction)*

Source: author's concept, design and calculations (Iliyan Petrov's PhD dissertation in Economics, 2015)

Experiments with different *"reference thresholds"* help to select an optimal position for the *inflection point - 0.5* for marking the transition to *50% majoritarian domination* and the value of "integral valuing information" is achieved in the ***variant SCIbas(4-1000),*** (table 2). The integral $\int_0^1 SCIbas(4-1000) = 0{,}4664$ is 41% higher than $\int_0^1 HHIbas = 0{,}33$ and 86% higher, than $\int_0^1 SWIbas = 0{,}25$ (in the case of normal logarithms *"ln"*). The distribution of information in *SCIbas(4-1000)* is better balanced in the intervals 0÷0.5÷1 and provides a "golden section", which is suitable application in large number of sectors (incl. energy, transport, machine building, etc.). For services sectors (banking, insurance, telecommunications) we may select other *SCI variants*, but retain ***SCIbas(4-1000)*** as reference model.

Table 2 - Inflection point and integral information in SCI, HHI and SWI models

Indicators/Models	SCIbas						HHIbas	SWIbas
Basic Model's variant	(2-1000) b_1=0,5 b_2=0,001	(3-1000) b_1=0,367 b_2=0,001	(4-1000) b_1=0,25 b_2=0,001	(5-1000) b_1=0,2 b_2=0,001	(6-1000) b_1=0,166 b_2=0,001	(10-1000) b_1=0,1 b_2=0,001	Single Variant s_i^2	$-s_i.\ln s_i$ $-s_i.\log_2 s_i$
Integral of structural information $\int_0^1 f(s)\,ds$	0,4226	0,4496	**0,4664**	0,4723	0,4798	0,4826	0,33..	0.25 0,36
Parameters:	Inflection point						No inflection, minimum, maximum	Maximum s | f(s)
Market share *(s)*	0,62	0,55	**0,5**	0,48	0,47	0,45		0,367|0.367
Basic function *f(s)*	0,54	0,49	**0,47**	0,44	0,435	0,43		0,367|0.53

Source: author's calculations

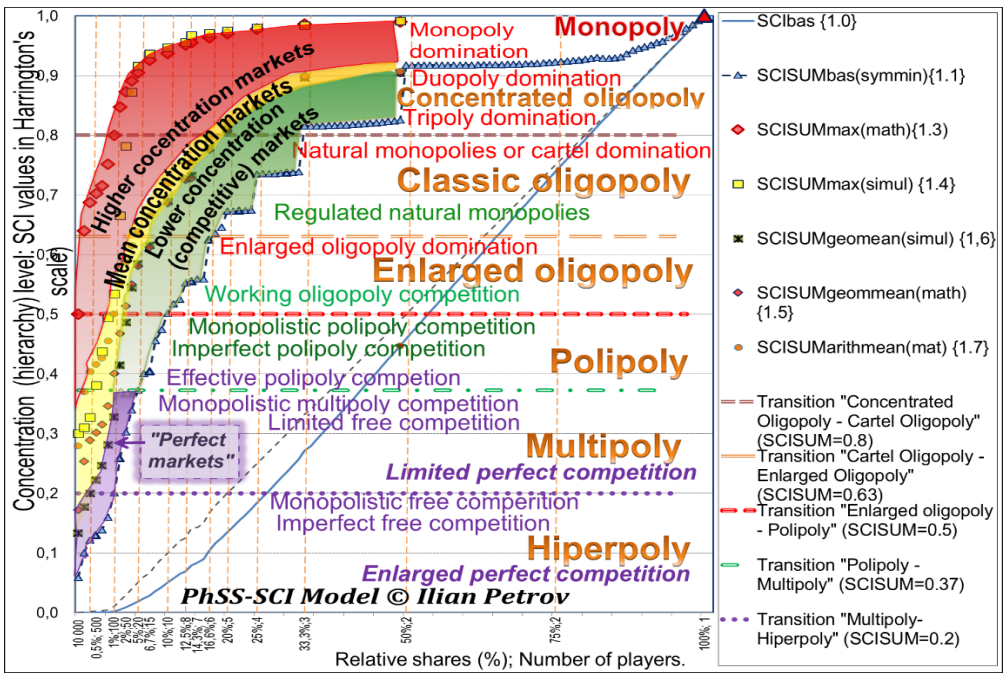


Figure 1 – Market structures development and competition interactions
Source: developed by author

The *PhSS-SCI model* does not assign separate stage for the abstract "perfect competition". However, the least concentrated *PhSS* in the stages of *polipoly, multipoly and hyperpoly* are referred as "free markets" with different levels of „perfect competition" (fig. 1).

Similar well-balanced evaluation scale and redundant classification is impossible even with sophisticated secondary manipulations in the other models (HHI, SWI, etc.). From this point of view, our SCI model provides more objective and balanced approach on *"structural information"* and *hierarchy* in different economic, social and political structures.

3 From Concentration to System Hierarchy and Coalition Synergy

Currently, for assessing market structures the economic theory employs the term "concentration", which is also used in antitrust legislation. Traditionally, this term is associated with a linear approach of measuring mixed states in physics and chemistry. However, its' mechanical transfer to economics creates some confusion. Multi-entity structures in economic systems should be examined from the point of view "hierarchy". Taking into account the role of tradition, it may take time for economic theory and legal systems to absorb innovative concepts. In accordance with common tradition our new indicator is, still, referred as *Set Concentration Index (SCI)*, but it should be perceived as a measure of hierarchy (*System Hierarchy Index - SHI).*

Table 3. Assessment and classification of structures, interactions and coalitions

SCISUM value	**Concentration level**	**Structure Stage**	**Competitive force and Market power**	**Value of CCC and Type of coalition**		**MTSI**
1	Maximum	Monopoly	Monopoly Oligopoly Domination	**Hyper cartel (CCC>0,8)**	Coalition "for itself"	Sellers Market Structure *MTSI>0,5*
0,8-1	Very high	Concentrated oligopoly				
0,63-0,8	High	Classic oligopoly	Classic oligopoly Cartel domination	**Super cartel 0,63<CCC<0,8**		
0,5-0,63	Moderately high	Larger oligopoly	Larger Oligopoly Monopolistic competition	**Cartel 0,5<CCC<0,63**		
<u>0,5</u>	**Zone of central qualitative transition boundary of structure concentration (hierarchy), "domination-competition", "cluster-cartel" coalition interaction and types of structures**					
0,37-0,5	Moderately low	Polipoly	Monopolistic competition Imperfect competition	Mini Cartel 0,37<CCC<0,5	Coalition "within it self	Buyers Market Structure MTSI<*0,5*
0,2-0,37	Low	Multipoly	Monopolistic & Free competition	Proto cartel 0,2<CCC<0,37		
0-0,2	Very low	Hyperpoly	Perfect competition	No cartel 0<CCC<0,2		
***<u>Abbreviations</u>*: SCI** – Set Concentration Index, **CCC**– Coalition Cooperativeness Coalition ; **MTSI** – Market Type Structure Index						

Source: author's calculations

For analyzing synergy, we introduce new indicator - the ***"Coalition Cooperativeness Coefficient" (CCC)***. It may include several criteria (with equal or different weights) for expert valuation, combining quantitative and qualitative assessment (tab. 3). The evaluation concept sets two zones divided by the central value ***CCC=0,5*** as a transitional boundary between less cooperative interaction ("cluster" or "coalition within itself") and more cooperative interaction ("cartel" or "coalition for itself"). The flexible concept of ***CCC*** may be applied in different economic, social and political sectors and reliable results may be attained with a limited set of criteria, covering 5 to 10 key aspects.

4 Market Structures and Interactions in the World Oil and Gas Sectors

This paper presents a simplified application of *CCC* for analyzing synergy effects of coalitions in the oil and gas sectors. A major challenge for further applied research would be to replace experts' view with more analytical indicators for assessment of real time dynamics. A multilayer approach would face problems like big data treatment, qualitative categorization and covariance of trends in interdependent processes.

Our earlier test application in the oil and gas sectors (Petrov, 2015) included 6 criteria with equal weights based on an individual expert evaluation with constant values for the whole period (tab. 4). On the supply side since mid-1970-s the increasing cooperativeness *(CCC=0.75)* of the **"Organization of Oil Exporting Countries" (OPEC)** can be classified as *"quasi cartel"*. The cooperation of the **"Forum of Gas Exporting Countries" (FGEC)** starts in the 2000-s and, so far, is in the transition zone between *"mini cartel"* and *"quasi cartel" (CCC=0.5)*.

Table 4. Coalition Cooperativeness Coefficient for oil and gas sector

Evaluation criteria of cooperation resources	OPEC	FGEC	OECD/oil	OECD/gas
Techno-economic advantages (volumes and costs of developing reserves and production)	0,9	0,8	0,15	0,35
Capacity flexibility for influencing supply/demand	0,8	0,6	0,5	0,35
Comprehensiveness of common policy	0,4	0,3	0,5	0,3
Imperativeness of rules (i.e. production and export quotas, volume of strategic reserves)	0,75	0,4	0,5	0,2
Market transparence and limits of pricing tolerance (wholesale, retail)	0,75	0,4	0,4	0,25
Logistics efficiency and security of supply chains (transport, storage, wholesale, retail)	0,9	0,5	0,35	0,25
Cumulative CCC assessment	**0,75**	**0,5**	**0,4**	**0,3**

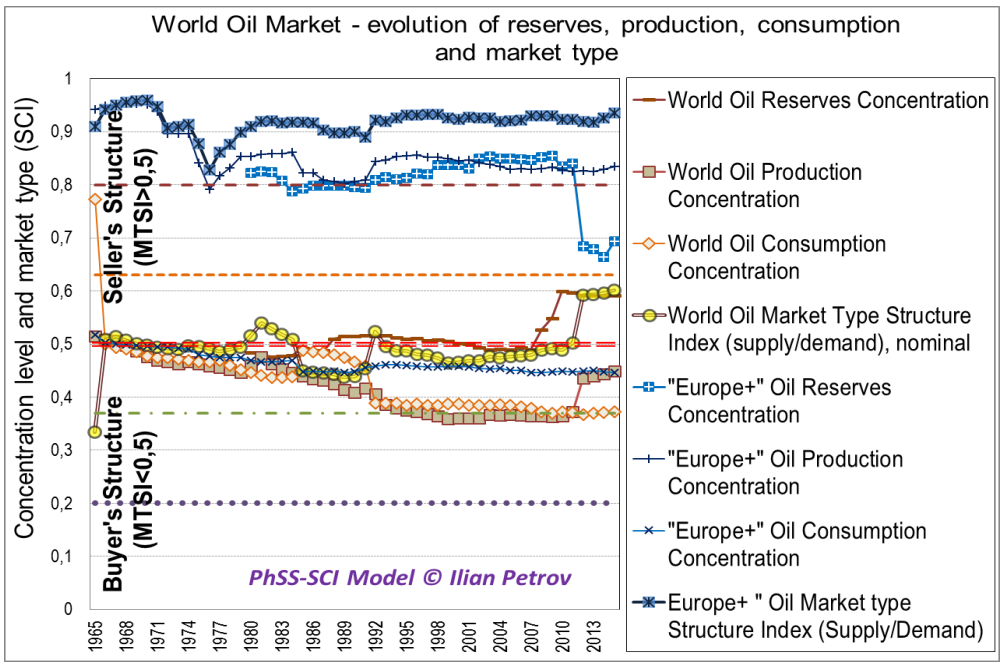
Source: author's individual assessment

On the demand side, the main developed countries are also key importers of hydrocarbons and since 1980-s their common energy policy within the **Organization of Economic of Cooperation and Development (OECD)** is coordinated by the **International Energy Agency (IEA)**. Traditionally the cooperation of OECD/IEA is more focused on the oil sector, where their effectiveness is higher *(CCC=0,4),* than in the gas sector *(CCC=0,3).*

4.1 Structural Evolution in World Oil Sector

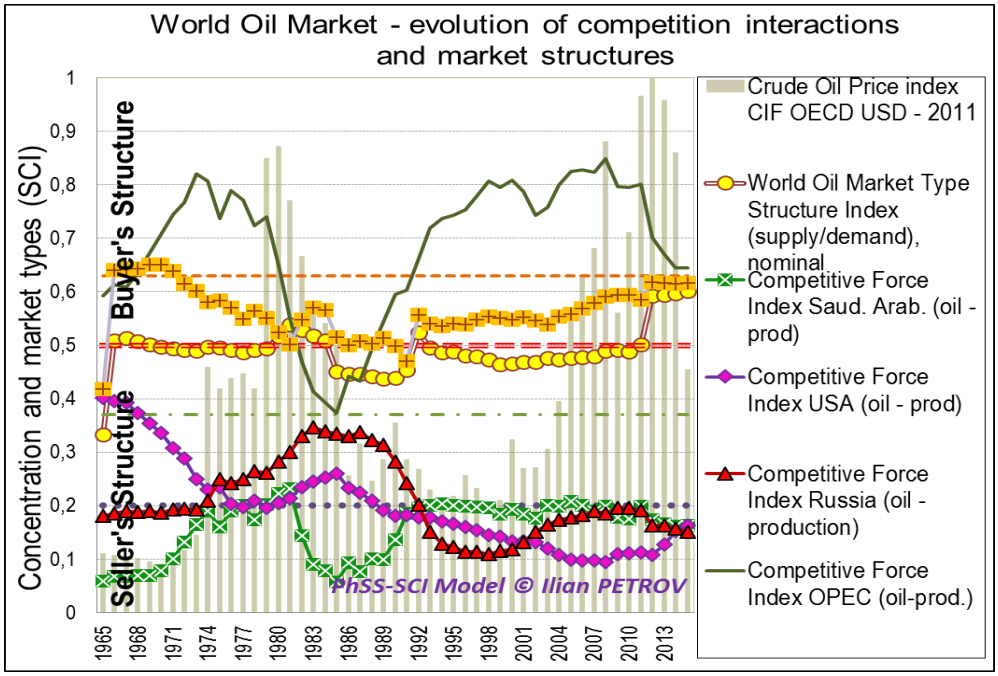
Our new *"Coalition Cooperativeness Coefficient" (CCC or "k")* facilitates the analysis of coalitions and their effects industrial structures, competition and price turbulences. In 1965-2014 we observe significant evolution in the structures in terms of concentration of production and consumption – from *"enlarged oligopoly" (0,5-0,55)* to *"polipoly-multipoly" (0,44-0,36).*

Demand and supply seem to have in general similar trends, which in fact have different dynamic parameters. As a result, in nominal terms (without taking into account the synergy effect of OPEC) the type of market structure evolves around the equilibrium zone (~0,5) with a shift from *„buyer's market"* in the period 1990- 2000 to *„seller's market"* in the period 2010-15.



Source: author's calculations

Saudi Arabia strengthen its' influence. *CCC* reveals the dynamics of OPEC' domination as a cartel on the export (supply) sided. In 1985-2013 the *"Competitive Force Index (CFI)"* of OPEC increases nearly 200% (CFI=0,4→ 0,8), while leadership of Saudi Arabia becomes stronger ~ 280% (CFI=0,06→ 0,16).



Source: author's calculations

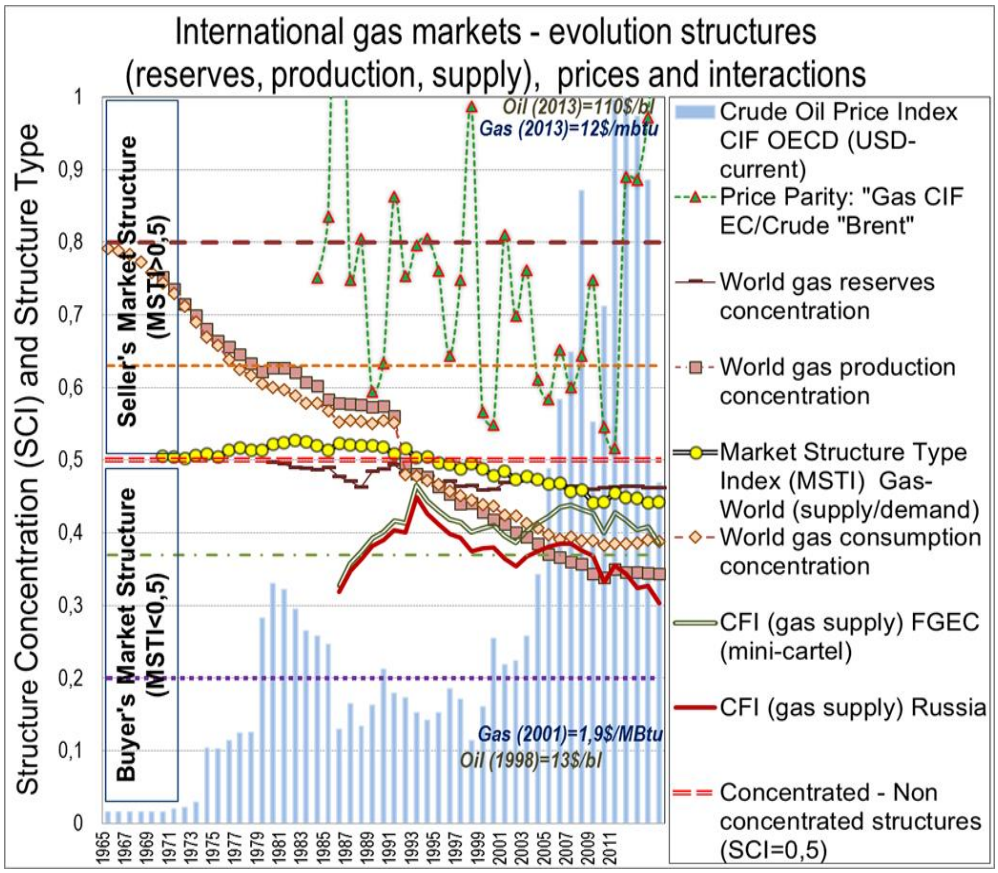
Russia's *"competition force"* is unstable – dramatic fall in 1980-1995 (CFI=0,36→ 0,14), recovery in 1995-2008 (CFI=0,14→ 0,2) and fall again since 2008 (CI=0,2→ 0,16). USA's *"competition force"* falls in 1965-2008 (CFI=0,43→ 0,10), but since 2008 recover (CFI=0,10→ 0,17), mainly due to improvement in "shale gas" technologies.

4.2 Basic Structural Trends in International Gas Markets

In the same fifty years period (1965-2015) the dynamics of industrial structure in the gas is more tangible than in the oil sector – the concentration of production concentration decreases faster (0,8-0,35) than that of Consumption (0,8-0,38). Improvement of production technology and lower concentration of world gas reserves (0,5 – 0,46) are basic determinant factors for such trend of diversification. As a result, Market Structure Type Index (MSTI) falls below 0,5 (0,53-0,46), shifting clearly from "*seller's"* to *"buyer's"* structure.

In terms of market interactions, the "*Competition Force Index (CFI)"* of the Forum of Gas Exporting Countries (FPEC) remains in the stage of *"mini cartel"* (CFI=0,4-0,44). If

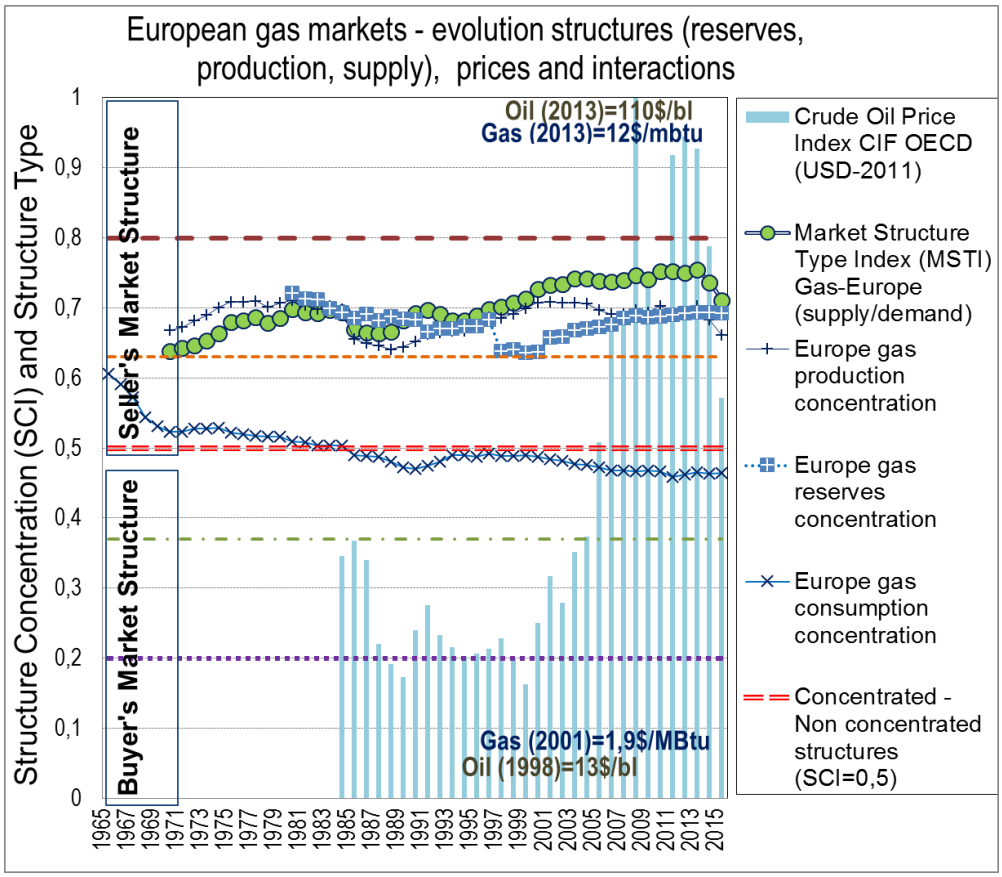
we take into account the major differences in vertical supply chains for "pipe gas" and "liquid natural gas (LNG), the cooperativeness of gas producers would look less active. These industrial and market aspects limit the trends of globalization and creation of a stronger "gas cartel" *(similar to OPEC in Oil Market)*, conserving the isolation of regional markets.



Source: author's calculations

Divergence and volatility of gas prices in different markets (long term contracts and gas hubs) is decreasing, while the price parity Gas/Oil still varies in large interval (1,2-0,5) with a stable lower limit.

Despite the attempts for unification and liberalization of gas markets the concentration of industrial structures in Europe remain high. Internal reserves and production decrease in volumes and their concentrations remain high (0,63→0,7). Innovation of technologies in the supply chain of LNG clearly contributes for decreasing the concentration of demand (0,61→0,47). The regional European market, being divided in very different national and sub-regional markets, has traditionally a "seller's" structure.



Source: author's calculations

Liberalization and diversification based on artificial increasing of number of intermediaries would hardly solve the problems of security of supplies and natural monopolies in Eastern Europe.

5 Conclusions

- Recapitulation of existing theories and clarification of some important notions confirmed the need for developing innovative concepts, methods and tools.
- Several original concepts, models and indicators are introduced and applied into practice: Phase Structure State (PhSS), Set Concentration Index (SCI) /or Set Hierarchy Index (SHI)/, Competition Force Index (CFI), Market Power Index (MPI), Market Structure Type Index (MSTI), Coalition Cooperativeness Coefficient (CCC or "k"). They are all integrated in a comprehensive "Structures Assessment and Classification Systems (SACS)".

- Improvement in existing concepts provide more logical and balanced system for classifying socio-economic structures into five main stages - „monopoly – oligopoly – polipoly – multipoly – hyperpoly“.
- The innovative PhSS-SCI methodology was successfully applied for analyzing structural evolution and market interactions in the world oil and gas sectors.

References

[1] Harrington, E., 1965. The disirability funtion. *Industrial Quality Control*, 21(10), pp. 494-498.

[2] Herfindahl, O. C., 1950. *Concentration in the U.S. Steel Industry.* Washington: Columbia University.

[3] Hirschman, A. O., 1945. *National power and the structure of foreign trade.* Berkeley: University of Califofrnia Press.

[4] Petrov, I. I., 2015. *Evolution of world energy markets structures and perspectives of development of gas infrastucture networks in South-East Europe (PhD Dissrtation).* Moscow: Gubkin Russian University for Oil and Gas.

[5] Prigozhin, I., 1990. *Time. Chaos. Quant..* Austin, TX (USA): University of Texas.

[6] Renyi, A., 1961. *On measures of information and entropy.* Berkeley, Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability 1960. pp. 547–561.

[7] Shannon, C., 1948. A mathematical theory of communication. *Bell System Technical Journal*, Volume 27 , pp. No.3 P. 379-423, No. 4. P= 623-656.

NARX Neural Network Application for Wood Resource Forecast

Pedja Milosavljević[1], Miroslav Milovanović[2], Milena Rajić[1], Dragan Pavlović[1]

[1]University of Niš, Faculty of Mechanical Engineering, Department of Management in Mechanical Engineering, e-mail: pedja@masfak.ni.ac.rs, milenatod1@yahoo.com, draganpavlovic10369@gmail.com

[2]University of Niš, Faculty of Electronic Engineering, Department of Control Systems, Aleksandra Medvedeva 14, 18000 Niš, Republic of Serbia, e-mail: miroslav.b.milovanovic@elfak.ni.ac.rs

Abstract: Wood resources planning and forecasting implies a challenging analysis. This analysis requires a complex data analysis which would take into account all inputs that define the yield of wooden material, which has direct impact on planning human resources, production timeline, as well as stock management of wooden assortments. This paper includes an analysis of monthly time series data from 1991 to 2015 which can be characterized as long time dependences data. In recent years, artificial neural networks have become a popular tool for time dependences data treatment. Therefore, a prediction of treated wood monthly requirements is performed by using the nonlinear autoregressive neural network with exogenous inputs (NARX model). NARX is a recurrent type of the network which is a very effective tool for approximation of any nonlinear function, especially ones which could occur during a nonlinear time sequence prediction. In the paper, different network topologies are tested and empirically is determined optimal structure which established as an excellent prediction tool in the field of wood science, engineering and technology.

Keywords: NARX neural network, Wood industry optimization, Forecasting, Resource management

1 Introduction

The real time processes could be highly non-linear and very unpredictable when some standard mathematical method is used as a prediction tool. Neural networks are successfully used in many analysis in business, industrial and scientific fields [1], in making models for transportation forecasting [2-3], in financial applications [4] and in the fields of energy production and operations [5-7]. Further, they are used in many prediction processes regarding wood industry and production of wooden assortments. Neural networks which are used to predict the future selling prices of fuel wood are presented in [8]. Combination of neural networks and ARIMA is used to forecast the number of cut Christmas trees in [9]. Identification of wood defects by the neural network is presented in [10]. In recent years, new

type of prediction network based on the artificial endocrine factor is developed [11, 12, 13, 14, 15]. This type of the network (endocrine neural network) is based on biological hormonal gland simulation. The artificial gland purpose is to stimulate network structure and make it extra sensitive to external factors. Implementation of endocrine factor inside a network effected network adaptation to environmental conditions in [11]. A great advantage of endocrine factor is that it can be implemented to various types of standard neural networks. A neural-endocrine system for realization of complex robot tasks is presented in [12]. It is shown that endocrine control logic can be very effective in predicting and calculating optimal trajectories, avoiding obstacles, picking up and dropping objects. Time series predictions are of main interest in our paper, and example for the realization of endocrine network for this purpose is presented [13]. Power management using the endocrine network is shown in [14] where is shown that this type of network can be very useful for managing and processing various types of data. In this paper will be presented a new type of endocrine neural network which is based on endocrine factor implementation inside traditional NARX (Nonlinear Autoregressive model with exogenous inputs) network, which will be presented within the next section. The proposed network will be used for forecasting market requirements of industrial wood quantities according to related production/sales history data set.

2 Neural Network Model

NARX model is a recurrent type of a neural network which is effective in modeling nonlinear system dynamics and time series forecasting procedures. Prediction capabilities of NARX are the main reasons for using this network as a base for further considerations in this paper. Basic NARX structure can be obtained from [16, 17]. Good forecasting performances of NARX are based on more effective learning procedures, faster convergence time and better generalization capabilities compared to traditional network performances.

Generalization capability is an important evaluation factor for choosing proper neural network structure when the task of a network is to forecast data. As it was said, NARX is a recurrent type of a network with improved generalization of long time dependencies [16]. Another important characteristic of this network is limited feedback architecture. Limited feedback indicates that feedback signal is based only on network output signals which are propagated back directly to the specified network inputs (Fig. 1). It is shown in practice that using NARX instead of default recurrent network does not involve larger network computation time. On the contrary, in the most cases, NARX network is reducing computation time required

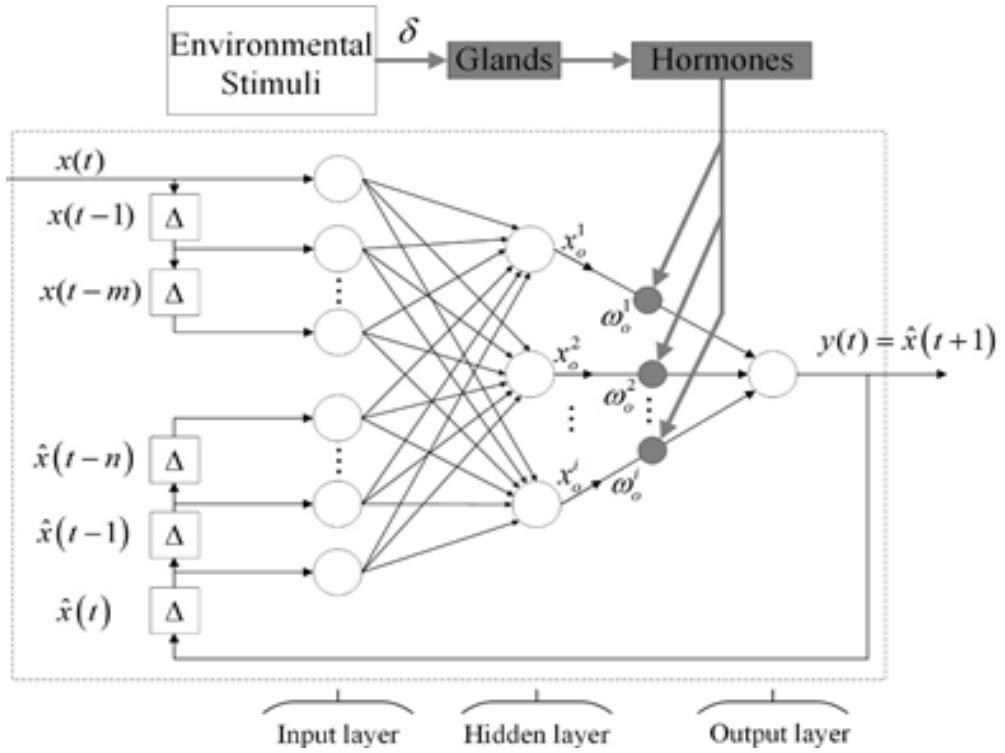


Figure 1. Endocrine NARX neural network architecture

for processing data. The structure of the NARX model can be presented as:

$$y(t) = f(x(t), \dots, x(t-m), \hat{x}(t-1), \dots, \hat{x}(t-n)) \quad (1)$$

where $x(t)$ and $\hat{x}(t)$ are input and output of the network at time t, respectively. Symbol f represents a nonlinear function which is approximated by multilayer network. Parameters m and n are the input and output memory order, respectively. NARX network outputs depend on m input past values and n output past values. Network input and output states are updating in each time point by principle described in [18, 19]. Additional information and characteristics of the default NARX network could be found in [18-20].

The proposed model in this paper is based on improved NARX model and inspired by the biological endocrine system. The biological endocrine system is used as a foundation for developing artificial endocrine systems in papers [11-15]. Important factors of biological systems are hormones which general purpose is to regulate various processes in a living being.

Hormone concentration which will be decayed by a gland is defined by gland stimulation level. This stimulation could depend on environmental influences, internal conditions and possible disturbances. Biological endocrine principles are utilized to control states and mimic adaptation processes inside artificial neural network. Artificial hormonal

glands are developed to mimic biological glands and they are implemented inside neural network structure to provide extra sensitivity and better adaptation to environmental conditions. The communication between environment and a network is performed by artificial stimulus δ (Fig. 1). The value of δ is a consequence of environmental conditions and it is making direct influence on implemented glands and provides network structure with additional environmental information. In papers [11-15] mathematical apparatus for developing endocrine network is presented. The most important part is to define the principles by which artificial gland will make influence on a neural network. Endocrine factor does not have influence on input and hidden layer neurons. The output values of these neurons are obtained by multiplying proper input signals with weight coefficients assigned to specific synapses, summing them into overall signal, and processing this final signal by selected activation function. Implementation of endocrine factor inside network is obtained by introducing proper hormone concentration only to specific output layer weights (Fig. 1). First, output layer neuron value (before processing by activation function) will be presented in the form:

$$y(t) = \sum_i \omega_0^i x_0^i(t) \quad (2)$$

where $x_0^i(t)$, are inputs of output layer neuron, and ω_0^i are proper weight coefficients for $i = 1,2, \dots, n$. Endocrine factor is introduced to the neuron by follows:

$$y(t) = \sum_i \omega_0^i x_0^i(t) \sum_g C_g(t) S_g^i \quad (3)$$

where S_g^i is sensitivity of ith synapse and $C_g(t)$ is hormone concentration at time t. Sensitivity is a gland parameter which defines hormone influence degree on a network and can take the value between 0 and 1. Hormone concentration which will be decayed by a gland, in each time step t, could be presented as:

$$C_g(t) = \beta_g C_g(t-1) + R_g(t) \quad (4)$$

where β_g is a decay constant rate and $R_g(t)$is stimulation value. The training process of the proposed neural network is based on two parts. The first part is an input/output updating procedure presented in [18, 19]. Second part is Levenberg-Marquardt learning method with adaptive momentum [21], which is used for updating the weight and bias values during the training process.

3 Case Study

This study presents forecasting performances of the pro-posed ENARX network. The task is to forecast monthly quantities of treated wood for the purpose of obtaining industrial wood material. Monthly training data consist of quantities of untreated and treated wood

throughout 1991 until 2015. Also, training data conclude sale of the assortments vector for the same time period. This vector will be used as environmental stimulus δ. That way we want to make dependence between quantities of sold wood products per month and raw round wood supplies on one side, and treated industrial wood quantities as final product on the other side. A total of 1800 data points were used for this study. All collected information are randomly divided into three subsets: 70% of the data was used for training, 15% for validation and 15% for testing the proposed network.

The ENARX structure is selected after preparation of the training data and it is constructed using MATLAB software package. Network structure possessed two inputs (coniferous and non-coniferous round woods quantities) and two outputs (treated industrial wood quantities). Forecasted parameters depend on row wood quantity values which are used in the last two months and which represent network input vectors. Additionally, forecasted parameters depend on past values of sold wood products which are introduced to the network in the form environmental stimulus δ.

The model structure is conducted by experimenting with a number of hidden layer neurons and a number of tapped delays. The optimum number of overall tapped delays in a neural network is analyzed in [22], and based on this study, the best value is in the range between 12 and 30. The most suitable performances are obtained for 2 tapped delays for each input and output, and according to Fig. 1: $m = n = 2$. Finally, the overall number of original and tapped delay inputs of ENARX network was 12. Input predictors applied to the network are current and last two months data points for each input/output vector.

Trial and error method is used to determine the number of neurons in the hidden layer. The trial process is repeated until it is not determined that 10 neurons in the hidden layer provide the best performances. Sigmoid activation function is used for implementation of all hidden layer neurons. Linear activation function is used for output layer neurons. Empirically is determined that learning rate should be equal to 0,005. The prediction performance of the training procedure is evaluated using the Mean Square Error (MSE) method. ENARX setting procedure finalizes with determination of endocrine parameters. Number of glands connected to stimulus δ is determined by trial, starting with one artificial gland and increasing the quantity by adding one in each attempt. The best performances were obtained with two implemented glands. Hormone and stimulation parameters (C_g and R_g) are determined according to the procedure described in [11]. Value of sensitivity parameter (S_g) is chosen to be equal to 1 for each gland. In that case, the largest network sensitivity degree to external factors is provided. For network structure used in this paper, which possesses 10 hidden layer

neurons, the number of output layer weights which will be adjusted by endocrine factor is also 10 (Fig. 1): $\omega_0^1, \omega_0^2, \ldots, \omega_0^{10}$. It is decided that each gland provokes the same number of weights, and according to this, each gland will make influence on 5 randomly selected weights. Finally, as it was said at the beginning of this section, environmental stimulus δ is selected to be sale of assortments data which is collected for an observed time period. The simulation results of presented network and results comparisons with the standard NARX network will be presented in the next section. Another example of forming procedure of ENARX is shown in [23].

4 Results

Forecasting results of two networks, ENARX and NARX, are presented in Fig.2 and Fig.3. The evaluation time period includes data points from January 1991 until December 2015. During this time sequence, 298 monthly data points are processed by two networks. The predicted values are then compared with target values which should be achieved by forecasting procedure. Prediction performances of the two networks are presented in Table 1, and they are obtained by examination of MSE results, computation times and number of iterations. It can be concluded from the table that ENARX network possesses smaller forecasting error compared to traditional NARX network. Deficiency of ENARX is larger computation time which can be explained by a complex endocrine structure of the network. Third, larger computation time is not related to a number of iterations and it can be seen that ENARX training process includes a smaller number of learning iterations compared to NARX. It implies that complex ENARX structure requires more computation time per each iteration. It can be concluded that the proposed ENARX structure is improved version of NARX network with improved forecasting accuracy. The main advantage is network sensitivity to environmental conditions which improves its adaptive level to specified external factors.

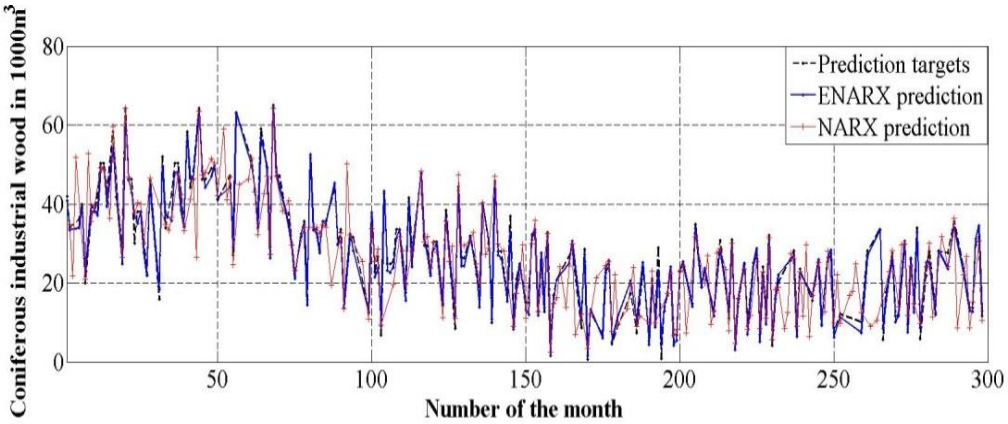


Figure 2. Coniferous industrial wood prediction–performances of two networks

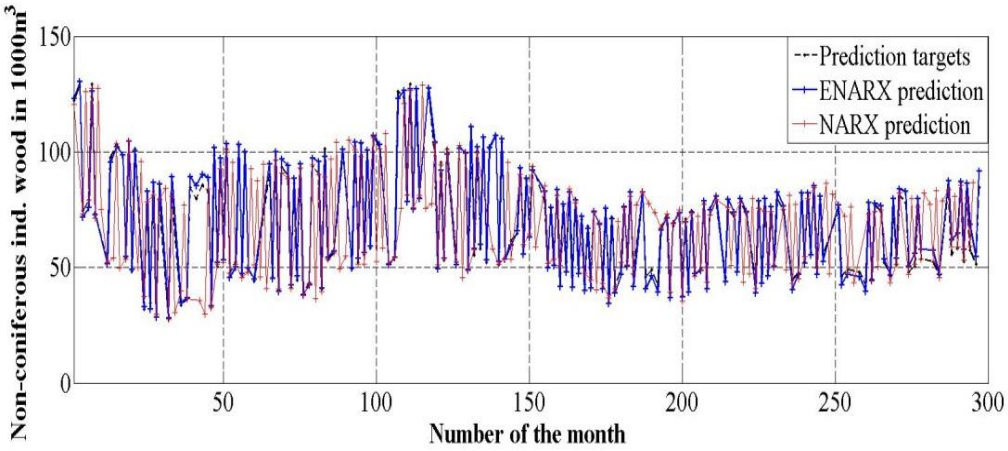


Figure 3. Non-coniferous industrial wood prediction–performances of two networks

Table 1. Evaluation of forecasting performances of two neural networks

	NARX neural network			ENARX neural network		
Evaluation parameter	**MSE**	**Computation time**	**Iterations**	**MSE**	**Computation time**	**Iterations**
Coniferous industrial wood	**6.12**	5.11 sec.	97	**0.884**	12.75 sec.	89
Non-coniferous industrial wood	**4.43**			**0.436**		

5 Conclusion

In this paper is implemented artificial endocrine factor in the form of artificial glands inside the standard NARX neural network. Implemented hormones are making influence only on output layer weights of the network. Leven-berg-Marquardt method is used for updating network parameters during the training procedure. Forecasting task of the network was to forecast monthly production quantities of industrial round wood for two types of trees: coniferous and non-coniferous. Monthly training data consisted of untreated and treated wood quantities of 25 past years. Environmental stimulus is presented to the network in the form of sale of assortments data. ENARX showed better forecasting performances compared to traditional NARX network, and proved as an excellent tool for processing various types of

data. ENARX training processes required a smaller number of iterations compared to traditional network requirements. Only deficiency is larger computation time, which is not significantly important measure if the forecasting accuracy is of the main interest in the research.

References

[1] Widrow B., Rumelhart D., Lehr M.A. (1994) Neural networks: Applications in industry, business and science. *Communications of the ACM*, 37, 3, 93–105.

[2] Dougherty M. (1995) A review of neural networks applied to transport. *Transportation Research Part C*, 3, 4, 247–260.

[3] Martinelli D.R., TengH. (1996) Optimization of railway operations using neural networks. *Transportation Research Part C: Emerging Technologies*, 4, 1, 33–49.

[4] Fadlalla A., Lin C.H. (2001) An analysis of the applications of neural networks in finance. *Interfaces*, 31, 4, 112–122.

[5] Azadeh A., Ghaderi S.F., Sohrabkhani S. (2008) Annual electricity consumption forecasting by neural network in high energy consuming industrial sectors, *Energy Conversion and Management*, 49, 8, 2272–2278.

[6] Hamzaçebi C. (2007) Forecasting of Turkey's net electricity energy consumption on sectoral bases, *Energy Policy*, 35, 3, 2009–2016.

[7] Geem Z.W., Roper W.E. (2009) Energy demand estimation of South Korea using artificial neural network, *Energy Policy*, 37, 10, 4049–4054.

[8] Koutroumanidis T., Konstantinos I., Arabatzis G. (2009) Predicting fuel wood prices in Greece with the use of ARIMA models, artificial neural networks and a hybrid ARIMA-ANN model, *Energy Policy*, 37, 9, 3627–3634.

[9] Konstantinos I., Arabatzis G., Koutroumanidis T., Apostolidis G. (2011) Forecasting of cut Christmas trees with Artificial Neural Networks, *Proceedings of the International Conference on Information and Communication Technologies for Sustainable Agri-production and Environment*, 8, 11, 507–518.

[10] Pham D.T., Soroka A.J., Ghanbarzadeh A., Koc E., Otri S., Packianather M. (2006) Optimizing Neural Networks for Identification of Wood Defects Using the Bees Algorithm, *IEEE International Conference on Industrial Informatics, Singapore*, 1346–1351.

[11] Milojković M., Antić D., Milovanović M., Nikolić S., Perić S., Almawlawe M. (2015) Modeling of Dynamic Systems Using Orthogonal Endocrine Adaptive Neuro-Fuzzy Inference Systems, *Journal of Dynamic Systems Measurement and Control*, 137, 9, doi: 10.1115/1.4030758.

[12] Timmis J., Murray L., Neal M. (2010) A Neural-endocrine Architecture for Foraging in Swarm Robotic Systems, *Studies in Computational Intelligence*, 284, 319–330.

[13] Chen D., Wang J., Zou F., Yuan W., Hou W. (2014) Time Series Prediction with Improved Neuro-endocrine Model, *Neural Computing and Applications*, 24, 6, 1465–1475.

[14] Sauze C., Neal M. (2013) Artificial Endocrine Controller for Power Management in Robotic Systems, *IEEE Transactions on Neural Networks and Learning Systems*, 24, 12, 1973–1985.

[15] Milovanović M., Antić D., Milojković M., Nikolić S., Perić S., Spasić M. (2016) Adaptive PID control based on orthogonal endocrine neural network, *Neural Networks*, 84, 80-90.

[16] Lin T., Horne B., Tino P., Giles C. (1996) Learning long-term dependencies in NARX recurrent neural networks, *IEEE Transactions on Neural Networks*, 7, 6, 1329–1338.

[17] Dzielinski A. (1999) Neural networks based NARX models in nonlinear adaptive control, *Proceedings of the International Joint Conference on Neural Networks*, 3, 2098–2103.

[18] Siegelmann H., Horne B., Giles C. (1997) Computational Capabilities of Recurrent NARX Neural Networks, *IEEE Transactions on Systems, Man, and Cybernetics—PART B: Cybernetics*, 27, 2, 208-215.

[19] Hatalis K., Pradhan P., Kishore S., Blum R., Lamadrid A. (2014) Multi-step Forecasting of Wave Power Using a Nonlinear Recurrent Neural Network, *IEEE PES General Meeting*, 1–5.

[20] Siegelmann H., Horne B., Giles C. (1997) Computational capabilities of recurrent NARX neural networks, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 27, 2, 208–215.

[21] Ampazis N., Perantonis S. (2000) Levenberg-Marquardt algorithm with adaptive momentum for the efficient training of feedforward networks, *International Joint Conference on Neural Networks*, 1, 126–131.

[22] Diaconescu E. (2008) The use of NARX Neural Networks to predict Chaotic Time Series, *WSEAS Transactions on Computer Research*, 3, 3, 182–191.

[23] Perić S., Antić D., Milovanović M., Mitić D., Milojković M., Nikolić S. (2016) Quasi-Sliding Mode Control with Orthogonal Endocrine Neural Network-Based Estimator Applied in Anti-Lock Braking System, *IEEE/ASME Transactions on Mechatronics*, 21, 2, 754–764.

Security Management of Organization's Data: A Data Leak Prevention (DLP) Approach

Ivan Gaidarski[1], Georgi Kutinchev[2], Pavlin Kutinchev[2]
[1]Institute of Information and Communication Technologies-BAS,
Acad. G. Bonchev Str. Bl. 2, 1113 Sofia, Bulgaria
[2] Infinity Ltd, Dragan Tsankov blvd 36, Interpred-WTC. 1040 Sofia, Bulgaria
i.gaidarski@isdip.bas.bg

Abstract: This paper provides an approach of managing data security in an organization with the help of Data Leak Prevention (DLP) solution. This approach gives a real picture of how the organization's data moves inside and outside the organization - channels, data format, users actions and etc.
DLP solutions can monitor, manage and control the data and the data channels. DLP makes it possible to identify sensitive information about the organization and manage the movement of this information inside and outside the organization. With DLP solutions can be found breaches in the organization's security policy and to identify the offenders. DLP solutions make it possible to prevent leakage of sensitive information, maintaining the competitive advantages of organizations, their know-how, patents, trade information, customer information, personal data and more.

Keywords: Big Data, Data in motion, Data in use, Data Leak Prevention, DLP, Device Control, Content Aware Control, Encryption, Mobile Device Management, MDM

1 Introduction

Information flow throughout the modern enterprise is the lifeblood of the organization. The need for users to be able to efficiently create information pertinent to their roles, while using services that don't disturb day-to-day workflow, with devices that may or may not be owned by your organization, is of paramount importance [1].

Today, businesses face many challenges in trying to protect sensitive information. Laptops store huge amounts of information, portable USB devices are becoming smaller and smaller and have the space to store entire databases and volumes of key business data. Nowadays, mobile devices regularly connect to both corporate and unsecured public networks, as well as a wide range of cloud-based storage solutions. This often jeopardizes the safety of data. Simply locking down users is no longer a viable solution to the problem. Access to

information is essential for companies, and globalization has transformed the way people collaborate. For example, a person located in USA can easily work for a company in Singapore thanks to collaboration tools. Without these tools, work in the 21st century may have serious setbacks [2].

2 The Problem

Employees can access or create content—some sensitive, some not—and send it outside the organization in a matter of seconds, often with little thought given to its contents. Confidential information included in internal documents and saved to unprotected, unknown or unmanaged repositories such as file systems, SharePoint or cloud repositories, can be widely dispersed, saved and shared without the knowledge of the originating party. Simply put, content including customer personally identifiable information (PII), can be inadvertently accessed and breached. Intellectual property (IP) can be misclassified and distributed. Relying on employee's awareness of what's considered sensitive information has proven to be a highly risky practice based on the continuing scores of lawsuits and regulatory actions. The obvious repositories of this sensitive and inappropriate material include file system storage such as Windows File Shares, SharePoint sites and cloud-based repositories. The problem arises when there is little central management over these repositories, and thus no visibility into areas of potential risk. Modern corporate enterprises create, receive, store and transport huge amounts of information every day. The average employee handles substantial volumes of information daily. Content created by individual employees in the form of reports, marketing plans, proprietary design documents, presentations and emails are often hybrid in nature meaning they contain pieces of content from many other information sources. As employees create, share and consume content, they often copy, paste and reference content from other sources. In many cases, only a small percentage of the information is actually fully reviewed and analyzed by the individual employee as to whether it contains corporate intellectual property or PII before they share it with others inside or outside the organization. In this way, inadvertent disclosures of IP and PII often happen [1].

What is data leak? In the information security context, "data leak" (or "data loss") means a security incident in which confidential, protected or sensitive bus ness-related data are accidentally or deliberately released to an untrusted environment or unauthorized users outside of the organization (external data leak) or become accessible to unauthorized users inside the organization (internal data leak).

What is leaked? It is any valuable business-related or personal information. It is not only payment card data stored in a bank or Protected Health Information (PHI) of patients kept by a hospital – for other businesses it may be corporate confidential information, intellectual property (IP), trade secrets, private customer data, Personally Identifiable Information (PII) of employees, or even state classified data [3].

How data get leaked? It happens when the organization does not really control who has access to its valuable business data and what is allowed to do with this data. Data breaches may be caused by external activities or internal reasons. External attacks usually involve malware infiltration through vulnerabilities in software used in the organization, as well as phishing and other social engineering techniques. Data leaks may have internal roots – including for instance, system glitches and misconfigurations. However, the major internal reason of data leaks in organizations is human nature. Employees, clients, contractors, partners – all legitimate users of the corporate IT system – are humans whose accidental mistakes, negligence, curiosity or misconduct may lead to data leaks. Some users may also fall victims of phishing attacks. Others may purposely steal corporate information.

What are the consequences? Once those people who are not supposed to see restricted access data get it in their possession, they may sell it, publish it or use it in other ways to negatively impact the organization. In any instance, data leaks may cause heavy financial and reputational damages, lead to large penalties, expensive litigations and loss of business, or even cause damage to national security [3].

Who is affected? The data leakage problem is not a marketing hype – it is happening every day across the world. Statistics from all credible sources shows the same threatening picture – data breaches affect businesses across all industries, as well as non-profit and government sectors. These are not only banks, chain and online retailers, insurance companies and hospitals. No one is immune to data breaches: just look at the indicative cases happened last year: tax authorities (U.S. Internal Revenue Service), military (U.S. Office of Personal Management), extra-marital affairs sites (Ashley Madison), and even a vendor that itself develops spyware (Hacking Team). Worse yet, this threat to corporate IT is growing: in the database of breaches maintained by Risk Based Security in cooperation with the Open Security Foundation, last year almost 4 thousand incidents have been registered worldwide, which is a 30% growth versus 2014 (Figure 1). The total number of compromised records in 2015 has reached 736 million [3].

The pandemic of data leakage and theft has already become a top-level challenge for corporate executive management – as vividly illustrated by the diagram (Figure 2) from the

latest Information Security Study conducted by 451 Research, one of the most professional analytical firms in the IT security industry. Executive managers have a strong material reason for such a concern – the data leak can cost a lot of money [3].

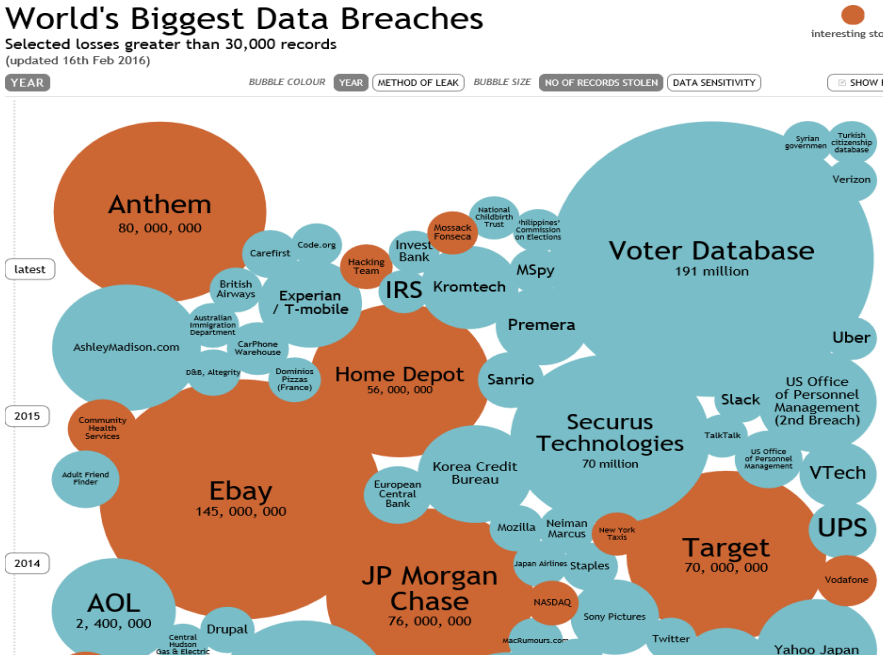


Figure 1: World's biggest data breaches [4]

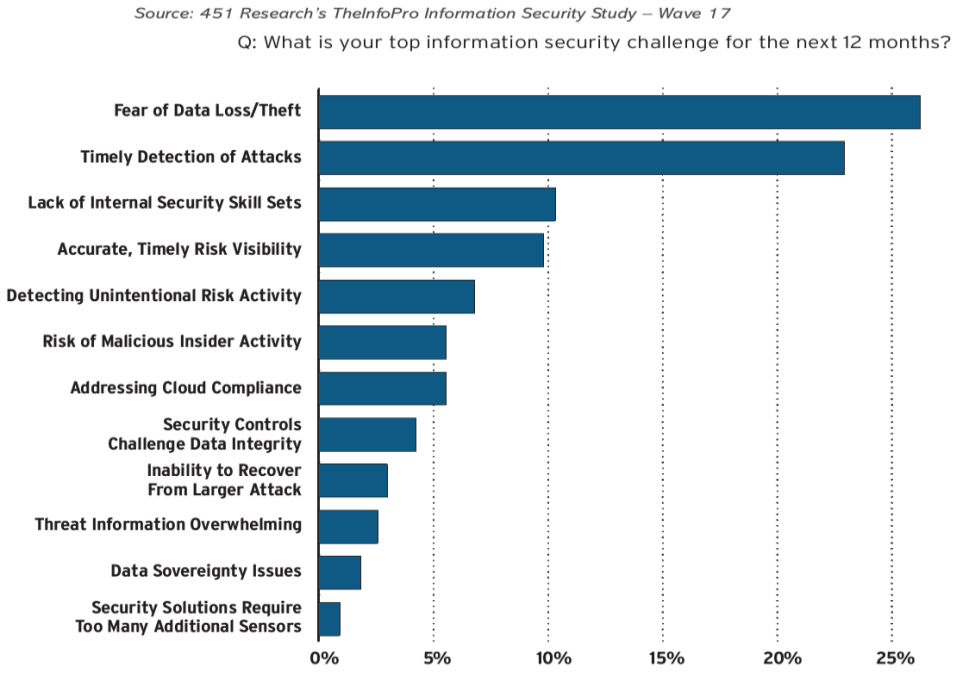


Figure 2: Top-level challenge for corporate executive management [5]

The Cost of Data Leaks (Figure 3). According to the latest study by the Ponemon Institute, on average in FY2016 financial losses of European companies for every compromised user record were in the range of $156 - $213 and in the U.S. they reached $221. The average total cost of a data breach for a business in EU was between $3.2 and $5 million depending on the country, while in the U.S. this figure grew to $7 million. However, in some specific cases – depending on the size and the business type – the financial damage to the company hit by the data breach can be two orders of magnitude more than the average. A real life example is a half billion class-action lawsuit filed against the owners of the AshleyMadison.com website followed a data breach in May 2015 that exposed millions of memberships in this adultery service [3].

Especially dangerous is that a significant part of all data leak incidents is related to insiders – normal users of corporate IT systems including employees, contractors, partners and clients. The reason is human nature – they make accidental mistakes, may be negligent, misconduct or become victims of social engineering attacks (e.g. phishing). The market and industry statistics shows that the majority of data leak cases involve insiders. Not only because they initiate incidents themselves but also because they make possible many of the externally initiated data breaches including such dangerous as social engineering and hacking.

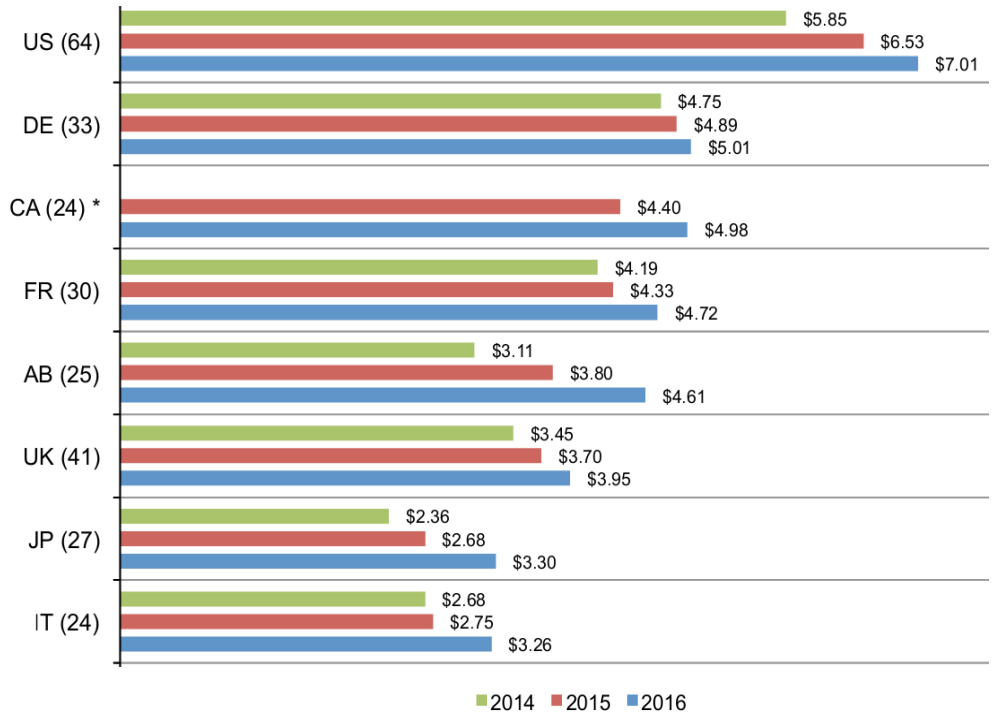


Figure 3: Cost of Data Breach Study, The Ponemon Institute, June 2016 [6]

In its 2016 global study, the Ponemon Institute estimates that a quarter of all data breaches in 2014 were caused directly by human errors. In addition, a significant part of another 48% attributed to the category called “malicious and criminal attacks” include incidents related to criminal insiders, as well as victims of phishing and social engineering. Therefore, it is safe to assume that the total percentage of insider-related data breaches accounted for in the Ponemon’s study exceeds more than 50% (Figure 4) [6].

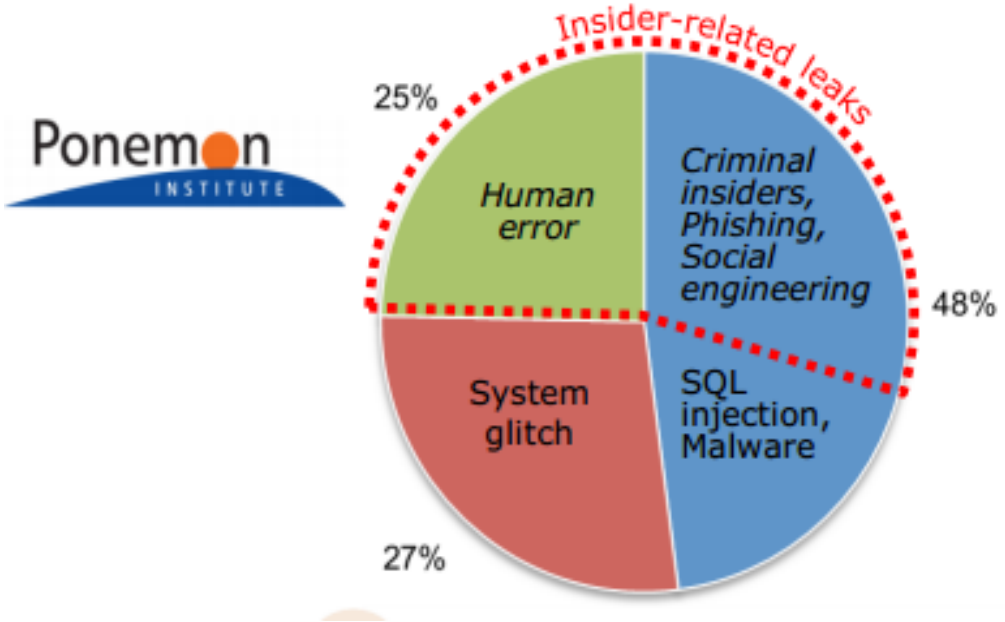


Figure 4: Insider-related leaks. The Ponemon Institute, June 2016 [6]

Even more critical dependence on insider actions in data breaches has been revealed in a report by BakerHostetler, one of the largest law firms in the U.S., in investigations of real data security incidents their customers had in 2014: in aggregate, 64% of all incidents were caused by humans including negligence (37%), insider theft (16%) and situations when users became victims of phishing (11%) (Figure 5) [7]. A dominating share of insider data leaks has been confirmed by a recent statistics from the Information Commissioner’s Office in the UK: of all data breaches reported to them from January to April 2016, 62% were caused by human errors. [8]

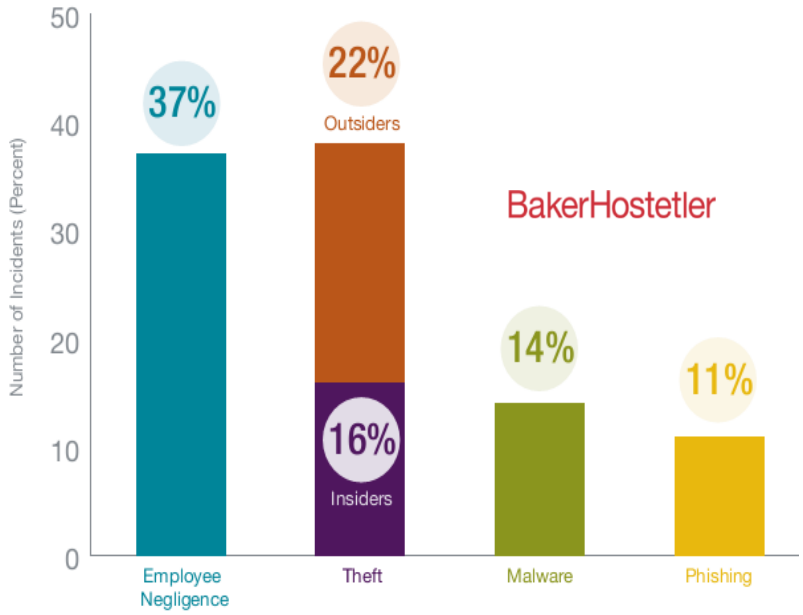


Figure 5: Data Security Incidents. Data Security Incident Response Report 2015, Baker & Hostetler [7]

3 The Concept of Data Loss Prevention (DLP)

The growing popularity of cloud storage solutions – along with ever-increasing compliance regulations designed to protect data – has made keeping confidential information safe inside the network more important than ever before. To protect sensitive information and prevent it from leaving the network, Data Loss Prevention technologies were developed to secure data while imposing minimal restrictions on users. As actions like sending an email or copying sensitive files to a USB device started to represent major sources of data breaches, DLP emerged to add a layer of security for which existing solutions were not designed. Unlike antivirus and firewall solutions, which focus on outside threats, Data Loss Prevention (DLP) targets inside threats; intentional or accidental.

Numerous definitions are acceptable to Data Loss Prevention experts, but a simple definition of DLP is: “Data loss/leakage prevention is a solution based on centralized policies designed to detect potential data breach by monitoring, detecting and blocking sensitive data”. A DLP solution inspects, allows or blocks data transfers to prevent data losses and thefts through various communication points like portable storage devices used with a desktop or a laptop, email applications, social media portals and cloud applications. DLP solutions prevent data loss from computers and laptops running various operating systems, portable storage devices and fleets of mobile devices (Figure 6).

Technology is changing fast, and DLP solutions are constantly adapting to protect data exposed by new ways of communicating, new file sharing tools and new devices. DLP technologies have evolved with the market. They do not impact employee productivity because they do not inhibit users from accessing helpful tools and services, unless the company chooses to block specific activities. Use DLP solutions that only restrict actions that could compromise confidential corporate data, and that easily adapt to and integrate with any type of network. Businesses can retain the latitude to define which information to treat as confidential in their environment, and whether to monitor transfers or block all unauthorized sharing of sensitive data. In sum, Data Loss Prevention solutions meet the multifaceted needs of today's business environment [2].

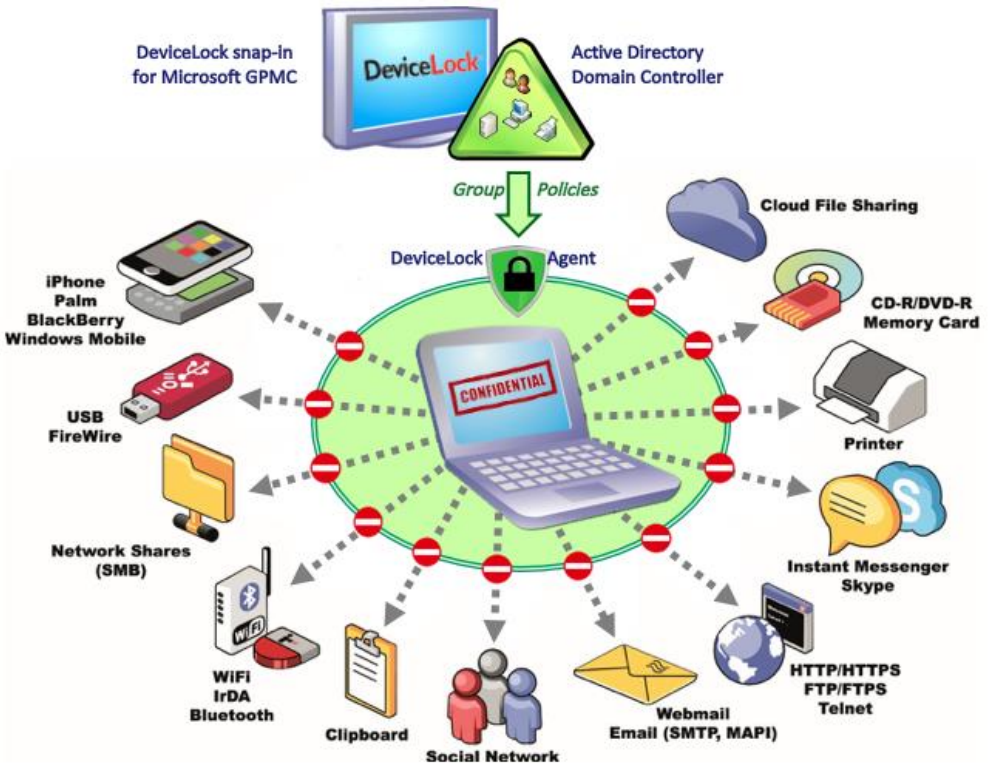


Figure 6: DLP solution by Device Lock [3]

4 DLP Technology Overview

Data Loss Prevention solutions are easy to use, yet very powerful tool that reduces the risk of accidental or intentional data loss. Without disrupting productivity, they enable a mobile workforce to take full advantage of portable storage devices, laptops, smartphones, tablets and cloud applications.

We will review the functionality of the DLP product “Endpoint Protector 4” developed by company CoSoSys (Figure 7). Endpoint Protector is an easy to use but highly effective cross-platform Data Loss Prevention solution. It was developed in response to the market’s need to stop data loss and data theft while continuing to take advantage of messengers, collaboration tools, social media, cloud applications and storage solutions. Endpoint Protector has a modular structure: Device Control, Content Aware Protection, Mobile Device Management and Easy Lock.

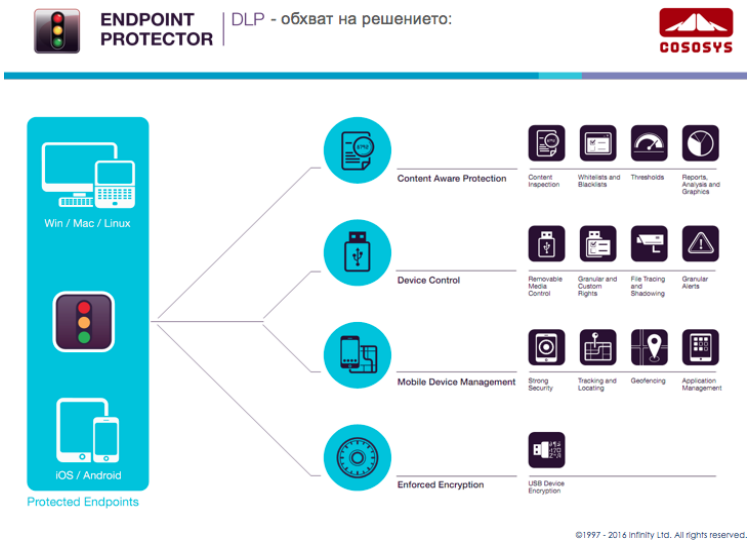


Figure 7: Endpoint Protector4 by Cososys

4.1 Device Control – Manage the Biggest Sources of Data Loss: Removable Media

There are many types of portable devices that connect to networks and have the potential to leak sensitive information. From CDs, DVDs, USB flash drives, memory cards, MP3 players, iPods, digital cameras and external HDDs to smartphones, everyone can easily store data on portable storage devices.

Device Control solutions monitor all the devices connected to computers and laptops and uniquely identify each of them. They include the option to set, Read Only Access’. ‚Allow’ or ‚Deny’ access to each device for each computer on the network, so managing USB and peripheral ports is easy. Based on each department’s activities and needs, a specific device can be authorized for use throughout the entire network or just on selected workstations (Figure 8).

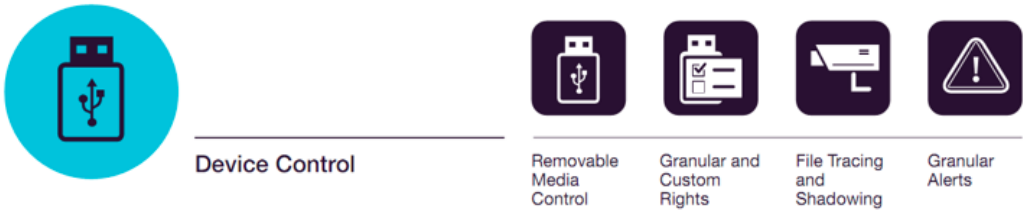


Figure 8: Endpoint Protector4 by Cososys. Module: Device Control

Advanced policies allow users to define different rights for a special category of devices from the same class, based on product ID (PID), vendor ID (VID) or serial number. These policies offer a powerful control and security function, especially for devices belonging to the same vendor from the same product line (same VID and PID, but a different serial number). The granularity of a Device Control solution is the defining factor between a flexible and powerful tool and a basic solution that provides limited options to manage rights and permissions. It is also the tool that supports Enforced Encryption on removable devices. It manages the usage of encrypted devices, be it hardware or software encryption that is used.

To secure data and encourage productivity, laptops and other devices must be able to exit the organizations. It is important to maintain assigned permissions for these laptops and devices when they're outside of the organization, and users can be given temporary access to a specific device or more devices for a limited period of time. The feature that enables this is called, 'Offline Temporary Password. It allows employees, if needed, to safely access devices and transfer documents like PowerPoint which may be needed at a meeting or a report during business travel. Furthermore, once the laptop reconnects to the company network, all logs are uploaded to the server, providing a complete report and overview of any attempted procedural violations. Such policies and tools are a must for today’s dynamic, mobile work environment.

File Tracing and File Shadowing are two other important features provided by an advanced Device Control solution. While setting up policies for all USB and peripheral ports is a great way to prevent leaks, it may not be adequate enough for some organizations. Access to detailed information and reports (size, file type, etc.) about attempts to copy data onto flash drives is also essential. In addition, the File Shadowing feature provides an exact copy of the transferred do ument, offering even more insight into which data left, or attempted to leave, the network. Part of Data Loss Prevention technology, Device Control represents the first line of protection against data loss and data theft. Besides blocking transfers of confidential data to

removable media, it also prevents malware proliferation, which is a leading cause of data corruption and data loss. [2]

4.2 Content-Aware Data Loss Prevention– In-depth Scanning of Sensitive Data

Content-Aware Data Loss Prevention enables the definition of more in-depth filters to determine whether information should be permitted to leave endpoints though portable storage devices, e- mails, web applications, cloud storage solutions and other means. Filters can be enabled based on file type, predefined content, custom dictionaries or, for those that have a deep understanding on the syntax, Regular Expressions. Enabled by just a click of a button, filters based on predefined content block documents containing sensitive information like credit card numbers (CCNs), social security numbers (SSNs) and other personal identifiable information (PII). Filters based on custom dictionaries extend the content aware protection to key words defined by each specific business entity, while filters based on Regular Expressions can cover a wide range of data (Figure 9). File Type Filter is also available for businesses that want to establish a wider policy and block transfer of documents based on their type.

Figure 9: Endpoint Protector4 by Cososys. Module: Content Aware Protection

As a second line of defense, Content-Aware DLP modules complement Device Control solutions, so transfers to removable media can be limited for certain documents and file types, and do not have to be completely blocked throughout the network. Since employees have different roles and varying needs for using specific data, protection should be granular and assigned per user, by computer, group and department, or made uniform for everyone throughout the entire company. There is also the option in the Content-Aware Protection DLP module to deactivate features like Print Screen or Copy/Paste if there are signs that people are using cloud applications to take data out of the company. Offline Temporary Password is available for Content Aware Protection as well. This means that mobile users that are not connected to the company network can receive a password for temporary rights to transfer confidential information through online apps. Again, productivity is not stopped for employees working during business travels or at home [2].

4.3 Mobile Device Management – Control the Smartest Devices

Not merely a consequence of IT consumerization or the BYOD trend, mobile devices are part of the corporate network and subsequently have confidential information residing on them. Mobile Device Management is a different concept from Data Loss Prevention and uses different technologies, but as part of best practices regarding data losses it should not be overlooked. With iOS and Android devices widely spread throughout any organization, an MDM solution has to protect information on both types of devices.

There are several important features that every MDM solution offers, like Tracking & Locating, Remote Lock or Remote Wipe. Using Remote Wipe must be done with caution because once triggered all data is deleted and the mobile device can no longer be managed remotely. Of course, MDM solutions provide additional helpful features like password-based policies, but a primary differentiation is a feature like Mobile Application Management (MAM) (Figure 10). While some vendors consider MAM a different solution entirely, remotely pushing or removing noncompliant apps is necessary within the MDM tool to have a successful mobile policy

Figure 10: Endpoint Protector4 by Cososys. Module: Mobile Device Management

Mobile Device Management solutions can increase productivity by facilitating the remote setup of WiFi, e-mail or VPN settings. However, it is important to remember that MDM is designed to prevent the loss or theft of mobile devices with company data stored on their SD card or on certain applications. When devices are lost or stolen, criminals often find it easy to hack into them because users most of the time create extremely simple passwords like "1234" or “password." Geofencing is the latest innovation for MDM solutions. It offers the possibility to apply policies depending on the location. Geofences can be created for the office building or for special delegations or branches. More context is added to the MDM policies, pushing certain settings or restricting the use of the mobile device camera within the desired perimeter, not in locations where it is not relevant [2].

4.4 Enforced Encryption – Harness the Trusted Device Technology

Another extension of a Data Loss Prevention solution is encryption. As mentioned, some users are given permission to copy data to USB storage devices, which can be lost or stolen once outside the company. By encrypting the data, third parties cannot access the information. There are numerous encryption solutions on the market, but one of the key differences between them is simplicity for the user. Easy to use encryption solutions offer simple Drag & Drop or copy/paste functionality, which automatically applies military strength AES 256bits encryption to data. Moreover, an encryption solution should seamlessly integrate with a Device Control solution, turning any USB stick into a Trusted Device and encrypting data in transit. If the encryption is tampered with, information is automatically deleted. This extra layer of security is an essential component of any Data Loss Prevention implementation (Figure 11).

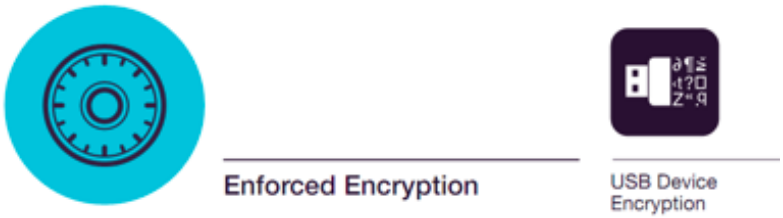


Figure 11: Endpoint Protector4 by Cososys. Module: EasyLock

The above mentioned features are not available in all the solutions on the market, and solutions that offer cross-platform enforced encryption are even more rare. Macs are becoming more popular in business environments, and although Apple provides built-in encryption, users tend not to utilize them. In addition, a USB flash drive is one of the most convenient ways to copy data to and from a Mac. Therefore, Enforced Encryption on removable storage for Mac OS X is a very good tool for preventing data loss [2].

5 Conclusions

As the number of internet-connected devices skyrockets into the billions, data loss prevention is an increasingly important part of any organization's ability to manage and protect critical and confidential information. Examples of critical and confidential data types include:

- Intellectual Property: source code, product design documents, process documentation, internal price lists.
- Corporate Data: Financial documents, strategic planning documents, due diligence research for mergers and acquisitions, employee information.

- Customer Data: Social Security numbers, credit card numbers, medical records, financial statements.

Data Loss Prevention has evolved to become much more than a solution for blocking USB ports or user access to data. It has grown into a complex and very granular solution that monitors, blocks or allows users to transfer confidential information based on policies determined within the company. DLP is not just about setting up roadblocks in a user's day-to-day activity, or being compliant with various regulations. It is about enabling users, providing access to the information, applications and tools they need to increase productivity, while also safeguarding sensitive data and preventing loss or leaks.

References

[1] Content is King - Controlling Content in Your enterprise – White Paper, Proofpoint, Inc., 2014

[2] Data-Loss-Prevention-Best-Practices , White Paper – CoSoSys Inc., 2014-2016

[3] DeviceLock_Intro for Customers_051016, White Paper, DeviceLock Inc. 2014-2016

[4] Informationisbeautiful.net , David McCandless © 2016; Risk Based Security, 2014-2016

[5] The Data Loss Prevention Market by the Numbers, 2014-2019, 451 Research, July 2015

[6] 2016 Cost of Data Breach Study: Global Analysis , The Ponemon Institute, June 2016

[7] Data Security Incident Response Report 2015, Baker & Hostetler LLP, 2015

[8] Press-release by Egress Software Technologies, 1st June 2016

Smart Place as a Service: A Model for Providing Big Data Solutions for Smart and Energy Efficient Buildings and Places

Roumen Nikolov; Alexandre Chikalanov; Elena Shoikova; Dimitar Paskalev;
Milan Rashevski;
r.nikolov@unibit.bg, a.chikalanov@unibit.bg, e.shoikova@unibit.bg, paskalev@izeb.eu,
mrashevski@gmail.com

Abstract: The paper is based on the outcomes of the SmartPlace project funded by a start-up grant from SpeedUP!Europe (speedupeurope.eu) – an accelerator under the EU FIWARE initiative (fiware.org). The main goal of the project is to develop and offer to the market open integrated solutions and services for management of inhabited places (apartments, houses, buildings, building complexes, open places) and analysis of Big Data. After the pilot solution for energy efficient and providing comfort and healthy environment for living "smart dormitory room", SmartPlace includes development of an integrated environment, which consists of three components: (1) integrated solutions for smart management of resources and comfort: using variety of smart sensors; appliances and other devices; registered useful models for lighting and heating; development of the technology infrastructure for data analytics, mobile applications, and user interfaces (Android/iOS); (2) a platform for sharing of: projects, algorithms, smart constructor with graphical interface for smart solutions dedicated to end-users (semi-professional/professional solutions) for management of smart places; (3) cloud infrastructure and services for Big Data analytics and support for providing a Smart Place as a Service (SPaaS) functionality. The envisaged products and services have innovative character in a global reach. There are a number of existing solutions for smart building management. However, most of them are closed in the framework of private technologies or standards of the product/service provider. There is no available analog of the platform for sharing solutions and the graphical constructor. The same holds for the SPaaS functionality. The main target group of the provided integrated solutions are the citizens – they can compose a creative, ever learning smart inhabited place by using visual programming and simulation tools in an open, intuitional virtual environment.

Keywords: big data, internet of things, smart home, smart city, energy efficiency, digital ecosystem, simulation, visual programming language

1 Introduction

The concept of Smart City as a means to enhance the quality of life of the citizen has been gaining increasing importance for technology providers, citizens and policy makers [1]. As an example, there are currently more than 100 cities from 23 countries in Europe, Latin America and Asia-Pacific which have joined the European Open & Agile Smart Cities (OASC) initiative that was launched on March 2015 (oascities.org). There is a clear link between the OASC initiative and the European Strategy for Digital Single Market. OASC promotes interoperability of systems based on the **free flow of data, between cities and within cities**, by adopting a shared set of simple, wide-spread, open and freely available mechanisms. This enables the development of better and more effective smart city applications and solutions which can reach many cities once created. The vision of the OASC initiative is to create an **open smart city market** based on the needs of cities and communities. OASC supports development of efficient services that **avoid dependence of a concrete vendor** and catalyse creation of a **city-based innovation ecosystem** oriented towards innovative SMEs and digital entrepreneurs. OASC boosts development of solutions based on **open innovation, open standards, open source and open data** and thus avoiding vendor locked-in solutions. Citizens participate in co-creation and experimentation processes for the development of technologies and services and thus achieving the necessary skills and competences needed to be citizens of a Smart City [2]. One of the technology platforms that are used as a backbone of the OASC initiative is FIWARE (www.fiware.org) which is the core of an independent open community whose members are committed to implement the FIWARE mission, namely "to build an open sustainable ecosystem around public, royalty-free and implementation-driven software platform standards that will ease the development of new Smart Applications in multiple sectors". The Smart Place project is based on the FIWARE platform and enablers and follows the above mentioned principles.

A Smart City must have suitable infrastructure, human capital and information. The digital dimension has a strong relationship with intelligence and innovativeness [3]. The linking component of Big Data and Smart Cities is the worldwide network of interconnected objects (Internet of Things) uniquely addressable based on standard communication protocols. The main characteristics of Smart Cities are divided into three forms of Intelligence:

- Orchestration intelligence: where cities establish institutions and community-based problem solving and collaborations;

- Empowerment intelligence: cities provide open platforms, experimental facilities and smart city infrastructure in order to cluster innovation in certain districts;
- Instrumentation intelligence: where city infrastructure is made smart through real time data collection, with analysis and predictive modeling across city districts.

The general objective of the Smart Place project is to develop a model and a Future Internet (FIWARE) based platform and services for energy efficiency and user comfort monitoring and management in a Smart City Building environment by using a combination of Big Data processing and simulation software.

2 Big Data and Simulations in Smart Home Solutions

There are a lot of research papers related to technologies and services in a smart home environment. A user-centred approach in the design and implementation of such solutions is dominating now. Nine research themes are identified and analysed in [4]. They are classified in three groups: (1) views of the smart home—functional, instrumental, socio-technical; (2) users and the use of the smart home—prospective users, interactions and decisions, using technologies in the home; and (3) challenges for implementation of the smart home - hardware and software, design and home-keeping. The smart home lets users to properly manage the inhabited place environment and resources and to improve the living experience through new functionalities, such as - remote control and automation of appliances, heating and lighting management, comfort management, security management, etc. In addition, smart homes are becoming nodes of a smart energy system that allows utilities to respond to real-time flows of information on energy demand fed back by smart meters from millions of homes [5].

A smart home may include: devices with a range of sensory capabilities such as programmable communicating thermostats, lighting and smart power strips; communications systems that facilitate two-way information flow between devices and the occupant, and possibly between devices and the utility or other third-party firms such as security system providers; and, monitoring and control systems that allow occupants to track energy usage and change the operations and functions of devices within the home [6]. The devices included in a smart home can vary tremendously in their design and function, but, in general, smart devices provide customers with the following two functions: ability to monitor energy use in real-time or near real-time for the whole house and/or by device; and, ability to remotely control systems or appliances in a home. The advent of smart homes may ensure smart technologies become a commonplace feature of people's lives, whether they are wanted or not [7].

There is a tremendous increase in the growth of data generated during the last decade. Society is in a process of digital transition and as a result, organizations are producing and storing vast amounts of data. Managing and gaining insights from the produced data is a challenge and a key to competitive advantage. Analytics solution as data mining can discover new patterns from large data sets.

"Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse [8]. Big Data refers also to various forms of large information sets that require special computational platforms in order to be analyzed. The design of big data systems keeps evolving when we need to handle larger-scale of data and more challenging user demands. Big data and the technologies associated with it can bring significant benefits to the business. But the tremendous uses of these technologies make difficult for an organization to strongly control these vast and heterogeneous collections of data to get further analyzed and investigated.

In July 2014, EC outlined a new strategy on Big Data, supporting and accelerating the transition towards a data-driven economy in Europe (https://ec.europa.eu/digital-single-market/en/towards-thriving-data-driven-economy). The data-driven economy will stimulate research and innovation on data while leading to more business opportunities and an increased availability of knowledge and capital, in particular for SMEs, across Europe.

Although a McKinsey report shows that the construction sector is still among the least "data intensive" sectors, Big Data is considered as the key factor for ensuring energy effciency in a smart building environment [9]. The energy performance model for buildings proposed by the CEN Standard EN15251[10] provides criteria for dimensioning the energy management of buildings, while indoor environmental requirements are maintained. It requires to identify the main drivers of energy use in buildings and after monitoring these parameters and analyzing the associated energy consumed, to model their impact on energy consumption, and then propose control strategies to save energy. During the monitoring phase, data from heterogeneous sources is collected and analyzed before specific actions are proposed to minimize energy consumption. Predictive models of the energy consumption of buildings could be created and used as well. Special attention could be paid on the behavior energy consumption of buildings due to the behavior of their occupants, the environmental conditions (temperature, humidity, pressure, natural lighting), and information about the energy generated in the building - alternative energy sources can be used to balance the energy consumption of the building. By analyzing the data generated one can identify any deviation between the predicted consumption and the real value. The system can provide behavior patterns of the

variables and develop models using such behavior patterns together with the associated sensed data. This approach helps implementing a predictive building model able to estimate the evolution of the energy consumption and to design strategies of control to save energy in the building based on the estimated evolution of the energy consumption. Once the energy usage profile of a building has been created, a set of appropriate actions to implement energy savings could be envisaged, e.g. strategies to adjust the operation time and configuration of the involved appliances or devices, selecting the optimal distribution of energy to maximize the use of alternate energies, etc.

Simulations and serious games are tools widely used in smart building applications[11]. IBM is among the leaders in Serious Gaming, e.g. INNOV8 (http://www-01.ibm.com/software/nz/solutions/soa/innov8/) is a simulation which brings IT and business together for process model innovation in three different scenarios: Smarter Traffic, Smarter Customer Service, and Smarter Supply Chains. CityOne (http://www-01.ibm.com/software/solutions/soa/innov8/cityone/index.jsp) aims to help urban planners, civic and business leaders to make cities "smarter" or more environmentally and socially sustainable. CityOne is considered among the top-10 serious games that changed the world. Most serious games contain fixed data and scenarios. However, IBM also produces games with real data flowing through them[12]. IBM utilizes its cloud infrastructure and applications, incorporate adaptive analytics (using the Watson system), business process management, learning management systems and social business applications. This provides a motivating, contextual environment for people to collaboratively solve problems. The SmartPlay Framework backed by artificial intelligence and the power of crowdsourcing – takes games to a new level (http://www-935.ibm.com/services/us/gbs/gaming/). These games contain real data, real business models and real data analytics.

3 The Smart Place Concept

The general objective of the Smart Place project is to develop a model and a Future Internet based platform and services for energy efficiency and user comfort monitoring and management in a Smart Building environment. Smart Place is an integrated system with great potential for market penetration, which consists of three main components:

a) Integrated solutions for smart management of resources and comfort of inhabited places (student dormitory, apartment, house, buildings, etc.): a variety of intelligent sensors and actuators, appliances; a set of appropriate models for lighting systems and heating; appropriate technology infrastructure and tools for Big Data analytics, mobile applications and

user interfaces (Android/iOS) – see Fig. 1 - 4. The model of **Fog Computing** is being used for analyzing and acting on IoT data. This allows to analyze the most time-sensitive data at the network edge, close to where it is generated instead of sending vast amounts of IoT data to the cloud. The system acts on IoT data in milliseconds and sends selected data to the cloud for historical analysis and longer-term storage [13].

The main outcome of this phase will be a well calibrated and evaluated (in a real environment with real users) prototype of a smart place (e.g. apartment), which includes a system for monitoring of energy consumption and comfort. This will lead to improved indicators for energy consumption and level of comfort (e.g. wellbeing). This prototype will be in the core of service for an integrated solution for a smart place, which (depending on the concrete needs and technology competence of the user) will include a complete product (software, sensors, actuators, controllers) with basic functionality, or – a service (design and implementation of an integrated solution, including a sensor network, appliances, etc.). The implemented smartphone application will allow monitoring the indicators of energy efficiency and comfort, providing control of the smart cyber-physical system, as well as to control different components. In addition, by using the concepts of Fog/Edge Computing and Cloud Computing, some functionality of data analytics and predictions will be implemented, e.g. behavior monitoring and predictions, creation of user profiles, etc. This will allow better personalization of the application based on the evaluation of the user experience. The application will use also instruments for learning, e.g. through serious games and gamification scenarios. Different award schemes will be promoted through specialized or general social networks aiming to stimulate users to reduce energy consumption and CO_2 emissions. The prototype is based on open source components from the FIWARE Platform and will provide services to individual users (see Fig. 1). These services will be available through a specialized Smart Place platform and dedicated to different types of inhabited places, such as apartments and homes. Remote management of the resources and comfort will be achieved by dedicated smartphone applications and specialized data analysis and recommendations for behavior change.

Smart Place as a Service counts on an intelligent self-learning platform, which is not just storing Big Data on the cloud, but also - analyzing behavior patterns and harnessing sophisticated algorithms for self-learning and optimization. The architecture is built according to a multi-tier paradigm (Fig.1).

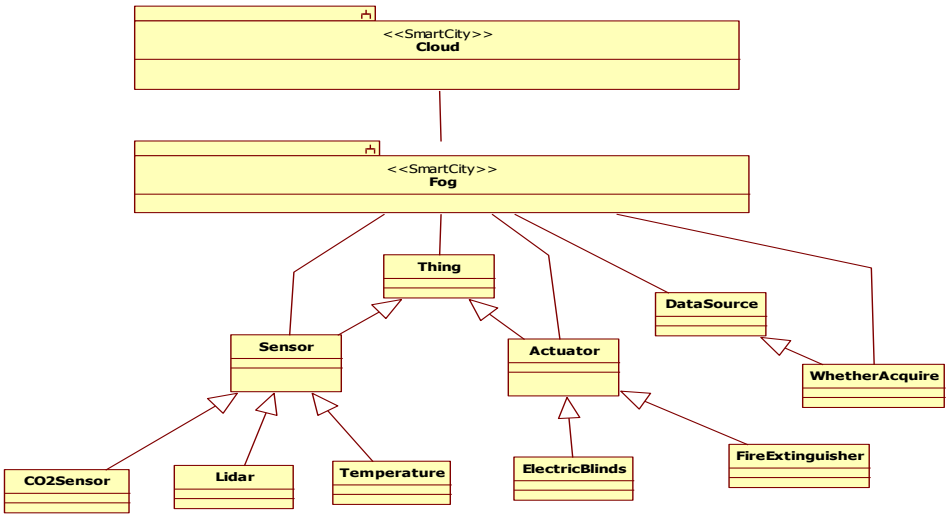


Fig. 1. High level multi-tier architecture of proposed application

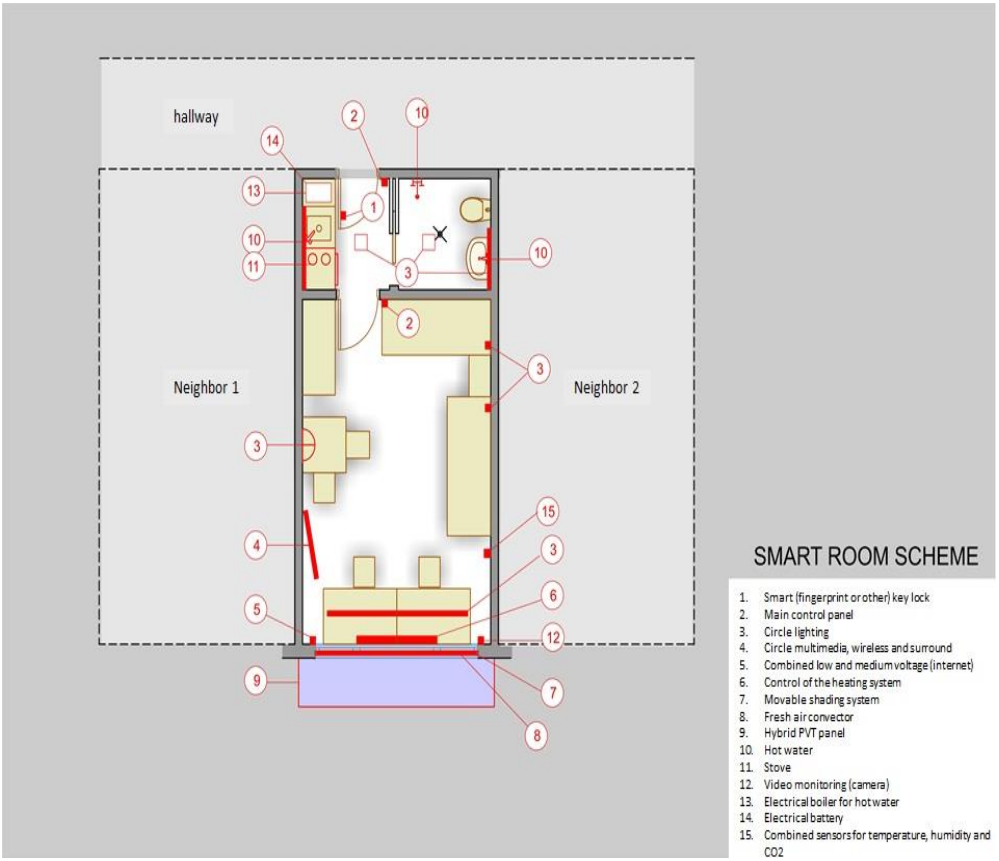


Fig.2. Smart Student Dormitory Room

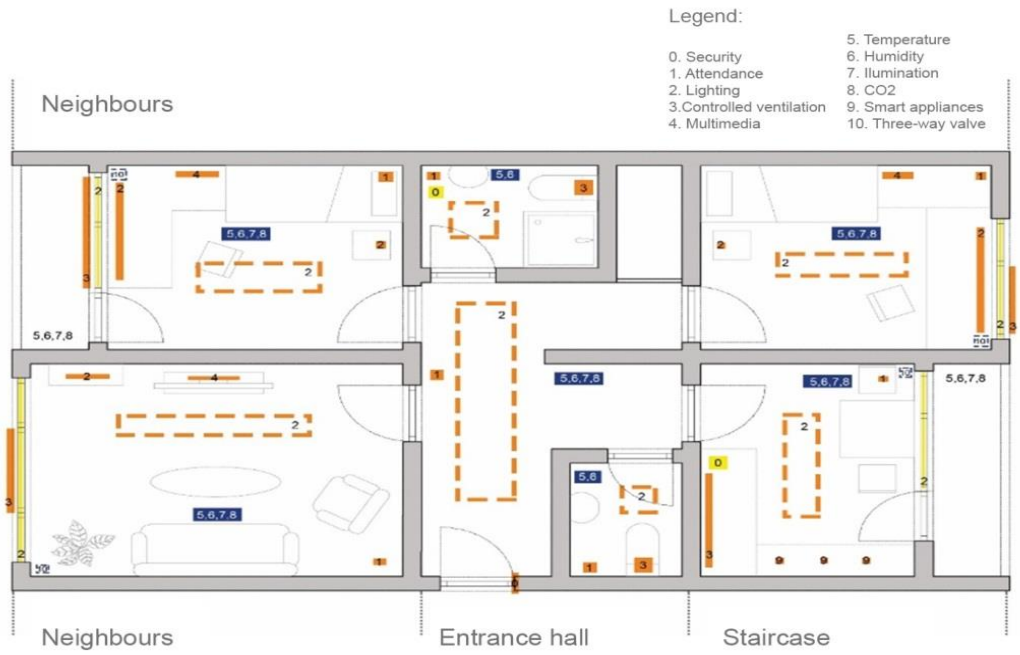


Fig. 3. Scheme of a typical apartment with sensors and appliances

Fig. 4. A prototype of a smartphone user interface

b) Open user-centred platform with graphical interface and visual language – a constructor (Lego type) of smart place solutions based on crowdsourcing and sharing of open-

source components and algorithms is envisaged. It could be used for design and implementation of smart inhabited places by different end-users (professional or semi-professional) and even – by ordinary citizens (Fig. 5). These solutions and constructed smart places could be used for creation of simulated cyber-physical environments and "what-if" experiments with real data, even – to generate data both by the physical and the simulated components. This approach resembles the CityOne simulation and SmartPlay Framework of IBM, but it allows the users to create their own simulation, serious games and gamification scenarios. The project will build a crowdsourcing model and a professional community that will share and contribute to further expansion of the Smart Place platform through open source components, simulated sensors, controllers (e.g. promoting the PLC as a Service model), simulated appliances, simulated sensor networks and ICT infrastructures, etc. The concept of the Smart Place as a Service (SPaaS) will be further elaborated as well, especially in the frames of the third phase of the project development.

The platform will be based also on other open source solutions, such as: SweetHome3D (www.sweethome3d.com/) – a free interior design application that helps drawing the plan of a house and arranging furniture on it and visit the results in 3D; Blender (www.blender.org/) – a free and open source 3D creation suite that supports 3D modeling, animation, simulation, compositing and motion tracking, and game creation; Unity 3D (www.unity3d.com/) – an open game development platform (game engine) supported by a large community of developers; Hyperfair (http://www.hyperfair.com/) - a 3D immersive virtual reality platform which provides data tracking and intelligence and allows to interact and collaborate with partners and customers as avatar and to showcase products and services.

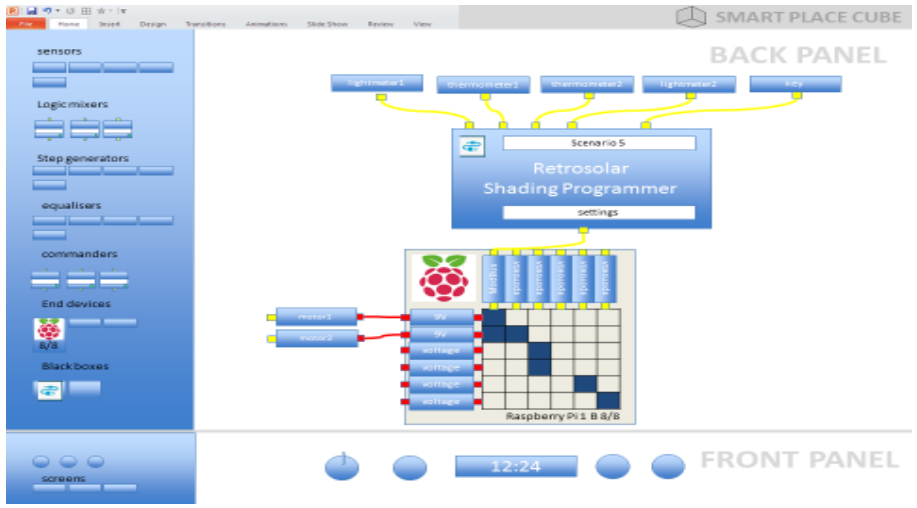


Fig. 5. The Smart Place Constructor

c) Smart Place as a Service Platform

The third phase of the project aims at defining the model of Place as a Service – SpaaS, and to develop a system prototype and services based on this model. This is in line with the tendency of providing models and cloud solutions of the type „Everything as as Service" (XaaS): software, platforms, infrastructure, business processes, mobile solutions, security solutions, etc. For instance: Mobility as a Service (https://maas.global/), Big-Data-as-a-Service (https://www.altiscale.com/), Hadoop-as-a-Service, Analytics-and-Visualization-as-a-Service, etc. Similar to the SPaaS model is the model Building as a Service - BaaS (www.baas-project.eu) explored in a FP7 Project. This model is dedicated mostly to commercial buildings. The SPaaS model is also related to a fast growing service sector – Place Management (http://www.placemanagement.org/) – which covers solutions in a range from building management to management of cities and regions.

The prototype of the SPaaS platform and services will be implemented on a Cloud Infrastructure by using Big Data analytics and visualization tools. The implementation will be based on the outcomes of the Phase 1 (already at a TRL 3 level - experimental proof of concept), the infrastructure built under the FP7 ELLIOT project (http://www.elliot-project.eu/, TRL 6 - technology demonstrated in relevant environment), and the outcomes of the Phase 2 (TRL 2 - – technology concept formulated). The SmartPlace software architecture (based on Fog Computing) will be extended towards full utilization of Cloud Computing by integration and adaptation of a number of FIWARE enablers, such as: Orion Context Broker; COSMOS; Cygnus; KeyRock; IBM Proactive Technology Online; EspR4FastData; Wirecloud; etc. The first prototype will provide a variety of use-cases, simulations and associated data related to the already implemented smart place solutions. At next steps the SPaaS platform will rely on Big Data sets provided by other building management systems (e.g. through the Big Data as a Service model) by implementing appropriate business model which takes into the consideration that **the data is the new global currency** (https://dupress.deloitte.com/dup-us-en/deloitte-review/issue-13/data-as-the-new-currency.html). Thus, the SPaaS platform could start providing global services in the construction and urban development sector that could be used by a variety of end-users (as co-designers, co-creators), including architects, designers, construction specialists, energy efficiency specialist, experts in public administration (national or municipal), producers of smart sensors, actuators and smart appliances, software developers, etc. The citizens in general could also be involved and they could build the necessary awareness and competences necessary for their behavior change and their support for adequate policies at national and municipal level. They could also experience personal

satisfaction of being part of campaigns for CO_2 reduction, energy efficiency, and ecology. The model of Living Labs (openlivinglabs.eu) will be used as well, which will allow building an appropriate innovation ecosystem around the SPaaS platform, that includes all stakeholders and stimulates digital entrepreneurship at regional and national level. The KBC methodology (Knowledge-Business-Social) for User Experience Evaluation developed in the frames of the ELLIOT project will be applied as well.

4 Conclusions

The main competitive advantage of the SmartPlace is its **holistic approach** - combination of both conventional construction technologies and solutions for energy efficiency using ICT (FIWARE, cloud and mobile technologies). The system is based on open source software and hardware, open standards and open data and thus avoiding vendor dependency of the SmartPlace solutions. It will use also a combination of Big Data processing and simulation software and thus - providing additional opportunities for offering flexible solutions ('what-if' functionality) for smart places, driven by: energy efficiency, budget, security and comfort. Smart Place is opening a new possibilities - to continuously learn and anticipate the needs and preferences of the inhabitants in terms of temperature, light and comfort.

Most of the home automation systems available on the market, are not based on open source software and hardware and they solve only custom cases through closed solutions. For instance, HomeSeer HS3 (www.homeseer.com) и CONTROL4 (www.control4.com) – the two top solutions on the market for 2016, do not use open source software and rely only on few standards (home-automation-systems-review.toptenreviews.com). The users experience difficulties in upgrading the provided solutions and in ensuring ongoing maintenance. Such systems could hardly provide data analytics based on multiple installations. However, the market leaders have agreements with a large number of producers of electronics and equipment and make efforts to ensure some good level of interoperability of smart city applications.

Other projects and initiatives based on open source software and open standards are Energy@home (www.energy-home.it) and Eclipse Smart Home (www.eclipse.org /smarthome). In April 2016 they announced their joint efforts to integrate their projects JEMMA and OpenHab. Prospective OpenHab solutions are the QIVICON platform of Deutsche Telekom (www.qivicon.com) and the ProSyst Bosch Group Smart Home platform (www.prosyst.com). Google Nest Labs and Apple HomeKit are not competitors, but rather technology providers that might help improving the SmartPlace solutions.

References

[1] Nikolov, R., Jekov, B., Mihaylova, P. (2015), Big Data in a Smart City Ecosystem: Models, Challenges and Trends, BdKCSE'2015 – Big Data, Knowledge and Control Systems Engineering, Sofia, 5-6 November, 2015, pp 111 – 118

[2] Nikolov, R., E.Shoikova, M. Krumova, E. Kovatcheva, V. Dimitrov and A.Shikalanov (2016), Learning in a Smart City Environment, Journal of Communication and Computer 13, pp 338-350

[3] Komninos, N (2011), Intelligent cities: Variable geometries of spatial intelligence, in: Deakin, M; Al Waer, H (2011), Journal of Intelligent Buildings International: From Intelligent Cities to Smart Cities, Volume 3, Issue 3, pp 172-188

[4] Wilson, C., Hargreaves, T., Hauxwell-Baldwin, R. (2015), Smart homes and their users: a systematic analysis and key challenges, Personal and Ubiquitous Computing, February 2015, Volume 19, Issue 2, pp 463–476

[5] Darby S (2010) Smart metering: what potential for householder engagement? Building Research and Information, 38(5), pp 442–457

[6] Saul-Rinaldi, K, LeBaron, R and Caracino, J (2014), Making Sense of the Smart Home. Applications of Smart Grid and Smart Home Technologies for the Home Performance Industry, National Home Performance Council, UK.

[7] Haines V, et al (2007) Probing user values in the home environment within a technology driven smart home project. Personal and Ubiquitous Computing 11, pp 349–359

[8] James Manyika, J., et al (2011), Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

[9] Moreno,M, et al (2015), Big data: the key to energy efficiency in smart buildings, Soft Computing, Springer-Verlag Berlin Heidelberg

[10] Nicol, F, Wilson, M (2010), An overview of the European Standard EN 15251, Proceedings of Conference: Adapting to Change: New Thinking on Comfort Cumberland Lodge, Windsor, UK, 9-11 April 2010. London: Network for Comfort and Energy Use in Buildings, http://nceub.org.uk

[11] Uskov A., B. Sekar (2015). Smart Gamification and Smart Serious Games, in: D. Sharma et al. (eds.), Fusion of Smart, Multimedia and Computer Gaming Technologies, Intelligent Systems Reference Library 84, Springer International Publishing SwitzerlandAmir Gandomi, Murtaza Haider. (2015) Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35 pp 137–144

[12] IBM (2011) IBM, Serious Solutions with Serious Games, https://www-935.ibm.com/services/multimedia/serious-games-overview.pdf, December 2011

[13] Cisco (2015), Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are, White Paper, https://www.cisco.com/c/dam/en_us/solutions/trends/iot/docs/computing-overview.pdf

Data Mining with Financial Open Data

Galia Novakova[1], Kamen Spassov[1], Silvia Popova[2]

[1] Faculty of Mathematics and Informatics, Sofia University

5 James Boutchier Str., 1164 Sofia, Bulgaria

[2] Institute of Systems Engineering and Robotics, Bulgarian Academy of Sciences

g.novak@fmi.uni-sofia.bg

Abstract: Platforms for open data are increasingly used in various fields of science, society and business, including urban planning, cultural heritage protection, crime prevention and more. Open data is becoming more prevalent and necessary and the need for effective and rapid analysis is increasing. Different methods of gathering and processing of open data provides data that the users can store and analyze according to their needs. Some requirements to open data are imposed as to be readable and regularly updated.

The aim of this paper is to gather and process big financial data in quarters for five year span and to test the proposed neural network model. Using this data and the knowledge extracted from the open data a classification of the selected companies and information about their development over the years is generated. Clustering analysis is applied to the problem of data mining because of its variety of applications. The chosen method for processing and aggregation of data is Neural Networks. This method is chosen because of its adaptability and efficiency. An implementation of the designed neural network is tested with financial statements of hundred companies. The result represents in most cases the expected predefined classification of all companies.[1]

Keywords: big data, financial statements, data mining, neural network, modeling

1 Introduction

The aim of the present paper is to gather, elaborate, generate and analyse knowledge from big financial data and to proposed a model for classification. Initially, data for quarters from financial statements of 100 companies in five year span is gathered. Every company listed

[1] **Journal of Economic Literature classification:** C1 (Econometric and statistical methods and methodology); D4 (Microeconomics - Market structure, Pricing and Design), L11 (Production, pricing and market structure; Size distribution of firms), L12 (Monopoly), L13 (Oligopoly and other imperfect markets), L4 (Antitrust issues and policies); L5 (Regulation and industry policy).

on the stock market is officially publishing its financial statements. There are many different sources of those financial statements. In the present paper we are checking whether the financial data of one company taken from one source is the same with the financial data proposed from other source. This check is important in order to have clear vision for information gathering and future development prediction. The analyzed 100 companies are classified into four categories: growing, stable, decreasing and others. The rest of companies will be defined between those classes aiming at testing the system. The companies' classification is important in order to understand what is the real market position of a given company and how it is developing.

Another aim of the paper is to use the initial classification based on financial data to train a neural network which ultimately could do this classification of companies by itself. Subsequently, the neural network will be tested with part of the companies for the whole period of time. It will generate results which will show for every single company the class of categorization.

The results from the neural network are shown for the whole period of development of a single company and this result is presented graphically for ease of reading and for the sake of visibility. Gathering the necessary financial information from one single source will assure the single record of financial statements and their fast analysis without being necessary additional processing. The results from the neural network can be offered to third parties which can perform by themselves the analysis and decide for subsequent actions.

There are four reasons for using neural network for those analyses:

1. The dependence between input and output data is nonlinear and the neural networks have ability to model non-linear patterns.

2. The neural network learns the main characteristics of a system through an iterative training process. It can also automatically update its learned knowledge on-line over time. This automatic learning facility makes a neural network based system inherently adaptive.

3. The neural networks make possible to define the relation (linear or nonlinear) among a number of variables without their appropriate knowledge.

4. There is a big number of data available. The neural network, trained with these data, adjusts the weights and predicts output with small error when working on new data with the same or similar characteristics of the input data.

The paper is structured as follows: Firstly, an introduction into the aims of the paper is made. Secondly, the financial data gathering and processing, and the proposed research methodology is described. Finally, in Section 3 are drawn some conclusions.

2 Financial Data Gathering and Processing

The companies that are selected in the present research are classified, according to the data obtained from their financial statements, in the following categories: growing, stable, decreasing and others. Bearing in mind this, first are gathered such companies that could be determined into those classes.

Growing companies are chosen by the criterion Most Innovative Companies [6]. After that, companies that are selected are checked in *Yahoo finance*[2] or in *Google finance*[3]. Those websites are free of charge and we could check the financial statements since the company is present on the stock market.

Market is dynamic, so the stock price is constantly up and down the reason being very different – from news for company's profits to news for deep recession. This is due to the fluctuation of demand and supply. The supply is changing with respect to the offered stocks on the market and the demand is changing with respect to the investors' will to buy or sell stocks.

In Fig. 1 demand is represented by the line having decreasing slope from left to right while supply is represented by line with increasing slope from right to left. The crossing point of the two lines shows the stock market price.

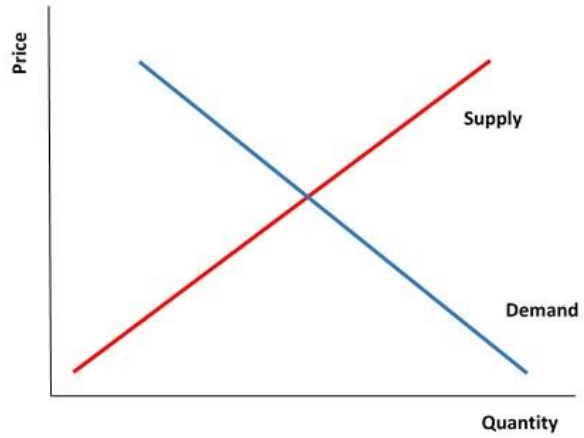


Figure 1: Optimal price is defined on the equilibrium (crossing point) between demand and supply

Demand and supply can fluctuate both with respect to the market situation. Different change of demand and supply can influence the share price [6]. Initially, the way one company could be classified in the category 'Growing company' is visually, looking at the share

[2] http://finance.yahoo.com/
[3] https://www.google.com/finance

movement. We choose for growing companies such companies that have a stable increasing graph for the last five years as shown in Fig.2.

Figure 2: Example of a growing company - Amazon

After the choice of the first 30 companies according to this criterion, we look for information in their financial statements. For choosing the data that is necessary to be gathered from the financial statements initially 40 factors are defined which best describe the financial results of one company and are the most important in the analysis. The information gathered for the selected companies is as follows: *Revenue, Total Revenue, Cost of Revenue – total, Gross profit, Selling/General/Admin. Expenses – total, Research & Development, Total Operating Expense, Income before tax, Income after tax, Net Income before Extra. Items, Net Income, Income Available to Common Excl. Extra Items, Income Available to Common Incl. Extra Items, Diluted Weighted Average Shares, Diluted EPS Excluding Extraordinary Items, Diluted Normalized EPS, Cash and Short Term Investments, Accounts Receivable – trade net, Total Receivables – net, Total Inventory, Other Current Assets – total, Total Current Assets, Total Assets, Accounts Payable, Notes Payable/Short Term Debt, Total Current Liabilities, Total Debt, Total Liabilities, Common Stock – total, Retained Earnings (Accumulated Deficit)), Total Equity, Total Liabilities & Shareholders' Equity, Net Income/Starting Line,*

Non-Cash Items, *Changes in Working Capital*, *Cash from Operating Activities*, *Cash from Financing Activities*, *Net Change in Cash* and *Status*. The company's status is defined as 1,2 and 3. Number 3 represents the companies which are growing, 2 stands for the stable companies and 1 is for decreasing companies.

For the rest of the categories, i.e. Stable and Decreasing companies, similar search is performed. We look for them according to the criterion „Worst client's service"[1]. The graphs of the stable companies do not have extreme ups and downs as could be observed in Fig.3.

Figure 3: Example of a stable company – AT&T

Some of the initially selected decreasing companies do not have financial statements for the five year span and those are of no interest for the purpose of the present paper. The missing data for a whole five year period makes the task of classification of those companies difficult and took more time (Fig. 4).

Finding the necessary open data as needed is not an easy task. Most of the web sites which offer such financial data are paid (see Fig.5) [2]. Some company's websites dispose financial data for such a long period however, the access to them is either not very easy or is under payment (Fig.6) [3].

Figure 4: Example of a decreasing company – Sears Holdings

Per Share Data	Annuals (USD $)																Quarterly				
Fiscal Period	Trend	Sep01	Sep02	Sep03	Sep04	Sep05	Sep06	Sep07	Sep08	Sep09	Sep10	Sep11	Sep12	Sep13	Sep14	Sep15	TTM	Mar15	Jun15	Sep15	Dec15	Mar16
Revenue per Share													23.65	26.21	29.86	40.34	40.17	9.94	8.59	8.89	13.56	9.12
EBITDA per Share													8.84	8.75	10.10	14.59	14.28	3.63	3.08	3.18	4.97	3.06
EBIT per Share													8.35	7.71	8.80	12.64	12.24	3.21	2.54	2.64	4.44	2.61
Earnings per Share (d...													6.31	5.68	6.45	9.22	8.99	2.33	1.85	1.96	3.28	1.90
eps without NRI													6.31	5.68	6.45	9.22	8.99	2.33	1.85	1.96	3.28	1.90
Owner Earnings per S...													6.88	7.13	8.28	12.00	9.91	11.27	12.13	12.00	11.24	9.91
Free Cashflow per Sh...													6.26	6.84	8.15	12.05	9.75	2.85	2.23	1.70	4.19	1.63
Dividends per Share													0.38	1.63	1.81	1.98	2.08	0.47	0.52	0.52	0.52	0.52
Book Value per Share													17.98	19.63	19.02	21.40	23.81	22.39	22.03	21.40	23.13	23.81
Tangible Book per Share													17.17	18.71	17.52	19.78	22.15	20.87	20.48	19.78	21.49	22.15
Total Debt per Share													–	2.69	6.02	11.56	14.58	7.61	9.54	11.56	11.38	14.58
Month End Stock Price													95.30	68.11	100.75	110.30	97.34	124.43	125.43	110.30	105.26	108.99
Ratios																						
Fiscal Period		Sep01	Sep02	Sep03	Sep04	Sep05	Sep06	Sep07	Sep08	Sep09	Sep10	Sep11	Sep12	Sep13	Sep14	Sep15	TTM	Mar15	Jun15	Sep15	Dec15	Mar16
Return on Equity %													42.84	30.64	33.61	46.25	40.05	43.02	33.54	36.32	59.32	32.52
Return on Assets %													28.54	19.34	18.01	20.45	17.80	20.75	15.99	15.79	25.16	14.06
Return on Capital - J...													475.62	313.84	289.44	339.94	317.03	369.52	284.24	280.61	444.02	254.28
Return on Invested C...													59.17	38.28	34.98	39.74	33.75	40.02	29.14	30.04	48.82	27.00
Weighted Average C...													8.11	6.94	10.05	9.75	11.86	8.75	10.59	9.75	9.65	11.74
Gross Margin %													43.87	37.62	38.59	40.06	39.81	40.78	39.68	39.90	40.10	39.40

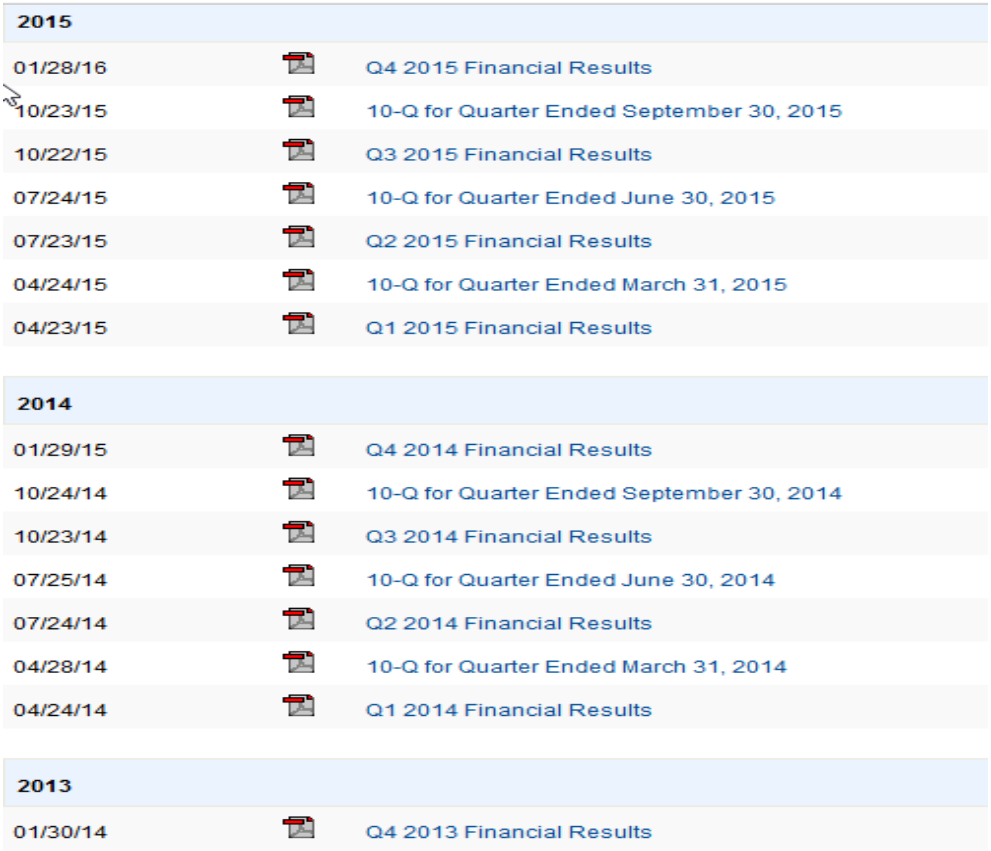
Subscribe to see/download 15-Year Financial data in one page.

7-Day Free Trial

Download 20Y Financial

Download 20Y Financial

Figure 5: Example of a paid website with data for many years ago



2015	
01/28/16	Q4 2015 Financial Results
10/23/15	10-Q for Quarter Ended September 30, 2015
10/22/15	Q3 2015 Financial Results
07/24/15	10-Q for Quarter Ended June 30, 2015
07/23/15	Q2 2015 Financial Results
04/24/15	10-Q for Quarter Ended March 31, 2015
04/23/15	Q1 2015 Financial Results

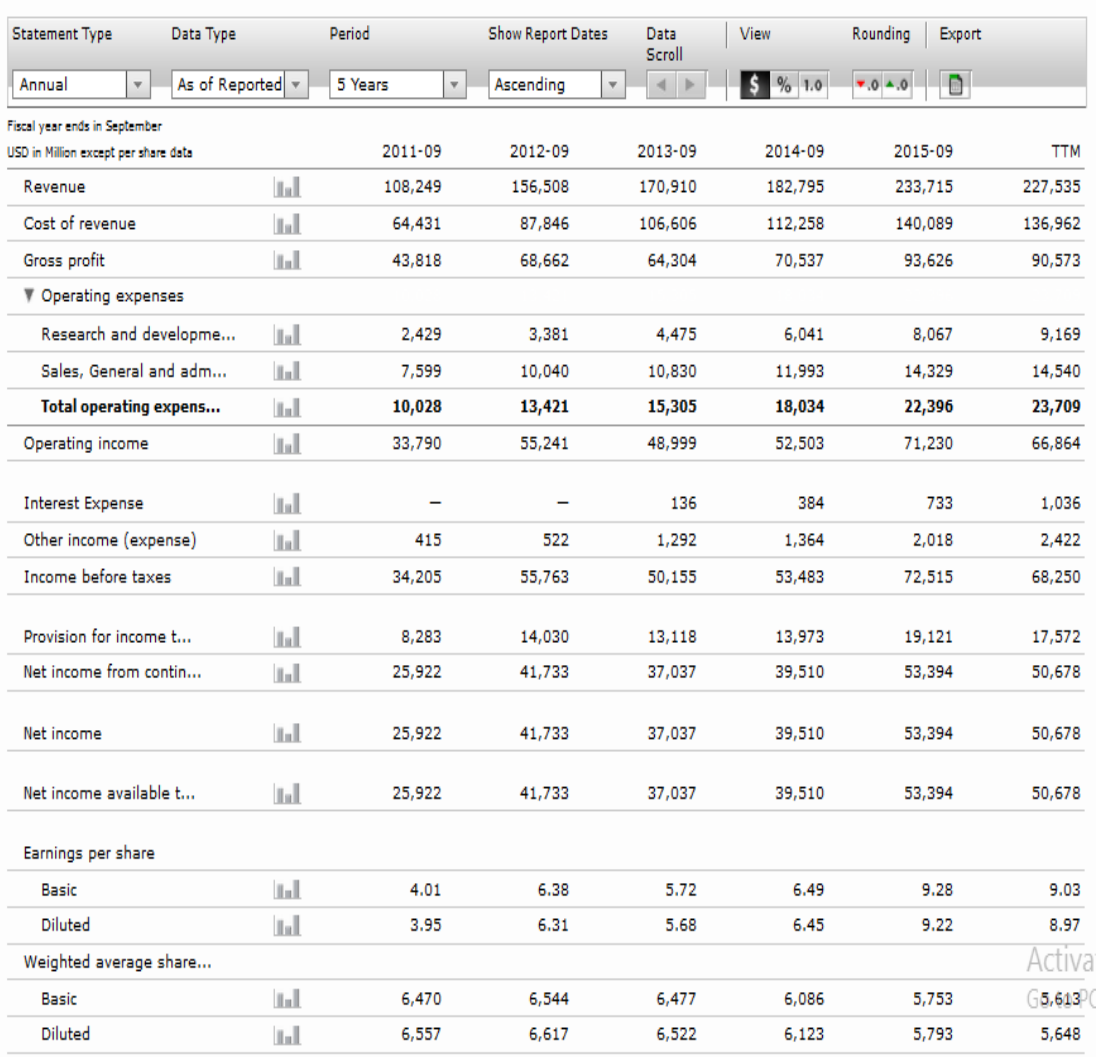
2014	
01/29/15	Q4 2014 Financial Results
10/24/14	10-Q for Quarter Ended September 30, 2014
10/23/14	Q3 2014 Financial Results
07/25/14	10-Q for Quarter Ended June 30, 2014
07/24/14	Q2 2014 Financial Results
04/28/14	10-Q for Quarter Ended March 31, 2014
04/24/14	Q1 2014 Financial Results

2013	
01/30/14	Q4 2013 Financial Results

Figure 6: Example of a company for which there is missing data

One of the obstacles for direct elaboration of open financial data is the fact that there is no given standard for their identical record. The extracted financial information usually is based on the financial statement information which on the other hand is often proposed in different format like in millions of US dollars, Korean yen, British pounds, etc. This different record of the necessary information ask for an additional step in extracting and analyzing the information, namely presenting it in one single payment system. In our case, the record that is chosen for the data in tables is in millions of US dollars.

Another problem is that some of the data sources give only part of the financial information for one company as shown in Fig.7 [4]. However, the authors have found a single source, which offers the complete financial statements for the five year time span, but that source is only 7 day free of charge [5**]**. All necessary data from the financial statements is structured in a table as shown in Fig. 8.



Statement Type	Data Type	Period	Show Report Dates	Data Scroll	View	Rounding	Export
Annual	As of Reported	5 Years	Ascending		$ % 1.0	.0 .0	

Fiscal year ends in September

USD in Million except per share data	2011-09	2012-09	2013-09	2014-09	2015-09	TTM
Revenue	108,249	156,508	170,910	182,795	233,715	227,535
Cost of revenue	64,431	87,846	106,606	112,258	140,089	136,962
Gross profit	43,818	68,662	64,304	70,537	93,626	90,573
▼ Operating expenses						
Research and developme...	2,429	3,381	4,475	6,041	8,067	9,169
Sales, General and adm...	7,599	10,040	10,830	11,993	14,329	14,540
Total operating expens...	**10,028**	**13,421**	**15,305**	**18,034**	**22,396**	**23,709**
Operating income	33,790	55,241	48,999	52,503	71,230	66,864
Interest Expense	–	–	136	384	733	1,036
Other income (expense)	415	522	1,292	1,364	2,018	2,422
Income before taxes	34,205	55,763	50,155	53,483	72,515	68,250
Provision for income t...	8,283	14,030	13,118	13,973	19,121	17,572
Net income from contin...	25,922	41,733	37,037	39,510	53,394	50,678
Net income	25,922	41,733	37,037	39,510	53,394	50,678
Net income available t...	25,922	41,733	37,037	39,510	53,394	50,678
Earnings per share						
Basic	4.01	6.38	5.72	6.49	9.28	9.03
Diluted	3.95	6.31	5.68	6.45	9.22	8.97
Weighted average share...						
Basic	6,470	6,544	6,477	6,086	5,753	5,613
Diluted	6,557	6,617	6,522	6,123	5,793	5,648

Figure 7: Example of a source which has only part of the data in the financial statements

Figure 8: Example of all necessary financial statements for one company – CarMax

3 Conclusions

The present paper deals with gathering and mining of financial open data for long list of companies. Once the financial statements of the companies are gathered in tables according to some attributes, an analysis is performed. Some changes in the initial structure of the data is done. The financial statements are used as an input data for the three-layer neural network with error backpropagation. A test file is created with randomly chosen companies. It has been chosen a neural network with 31 neurons in the inner layer which gives less error. The outcome of the proposed neural network with financial statements of the companies as an input data, is the 'companies status'. The status of companies is one of the following: Growing/ Stable/ Decreasing as well as a number which is given from the neural network and the status defined at the beginning. The neural network is giving as a result the company's name and its status. The table for analysis contains company's quarter data for five year span. The development and categorization of the companies could be presented graphically which gives better visualization and quick understanding of the company's development. The results and conclusions drawn for the neural network can be used for further development.

So that, the present paper gives ways for finding a reliable and complete financial data for companies as well as an analysis of financial statements and a choice for classification of companies. What can be improved is the way the financial data is gathered. This process can be automated with a macros or with a program that extracts data faster and easier from open sources.

The collected accounting data serving as an input data to the neural network is sufficient to ensure good training of the neural network and correct result. The neural network adjusts the weights during training so that, this number corresponds to the strength of belonging to the class. That numbers for the last three quarters give companys's development monitoring.

The results from the neural network at this stage will be used furtheron for:

- improving the design of the neural network to make a better forecast of the companies' development. Those companies forecasts can be used by investors for decision making;

- some of the paid financial sources of information which provide financial statements of companies could also be checked;

The neural network at this stage is trained with data only for known companies, for which we have information. It could be applied also for classifying companies which are relatively new on the market and need investing for future development. Another benefit of the proposed neural network is that it could be analysed if a given company is worth investing or not, and what is its status: Growing, Stable or Decreasing.

References

[1] DARPA Neural Network Study. October, 1987-February, 1989. MIT Lincoln Lab.

[2] Fahlman, S., Lebiere, C. (1990). The cascade-correlation learning architecture. *In Advances in neural information processing systems*.

[3] Gurney, K. (1997). An Introduction to neural networks (1997). London.

[4] Jones, A.J. (1993). Genetic algorithms and their applications to the design of neural networks. *Neural Computing and Applications* 1, 32–45 Print.

[5] Stergio, C., Siganos, D. (1889). Neural Networks in 1989.

[6] The Boston Consulting Group. (2014). Most Innovative Companies in 2014.

Simple Wireless Stack, Based on IEEE 802.15.4, Used for Process-control Applications.

Dichko Bachvarov*, Ani Boneva*, Yordanka Boneva*, Simeon Angelov**

* Institute of Information and Communication Technologies - BAS,

Acad. G. Bonchev Str., Bl. 2, Sofia 1113, Bulgaria,

dichko1952@abv.bg, a_boneva1964@abv.bg, dani_boneva@yahoo.com

**OMNITEL Ltd., Tsarigradsko Shousse blvd.125, bl.2, Sofia 1113, Bulgaria

simeon@omnitel.bg

Abstract: The article presents a simple wireless networking stack (based on IEEE 802.15.4). This stack is the extension of the popular IEEE 802.15.4, and provides developers design capabilities of managing wireless networks, bypassing the complexity of software structures typical of other stacks. The structure of the stack meet the requirements of process- control and design of multi-network-specific industrial applications. Included are options such as "routing" and "tree- topology" not supported by IEEE 802.15.4 (typical for large stacks). In the stack is implemented logical "multi-wire" processing, which enables the construction of multi wireless networks.

Keywords: IEEE 802.15.4, SWNS, process control, threads, topology, routing.

1 Introduction

This stack called SWNS is based on IEEE 802.15.4 and appears its extension. When developing SWNS are used some of the capabilities of IEEE 802.15.4, such as:

- The application uses functions of the 802.15.4 Stack API to interact with the IEEE 802.15.4 stack layers. This interaction is implemented in terms of MCPS/MLME requests and confirmations, indications and responses. The IEEE 802.15.4 stack interacts with the underlying hardware to access hardware registers.
- The application interacts with the on-chip hardware peripherals using functions of the Integrated Peripherals API. This API uses the peripheral hardware drivers to access hardware registers.
- The application interacts with the board hardware peripherals using functions of the Board API. The Board API uses the Integrated Peripherals API to achieve the interaction with the board hardware.

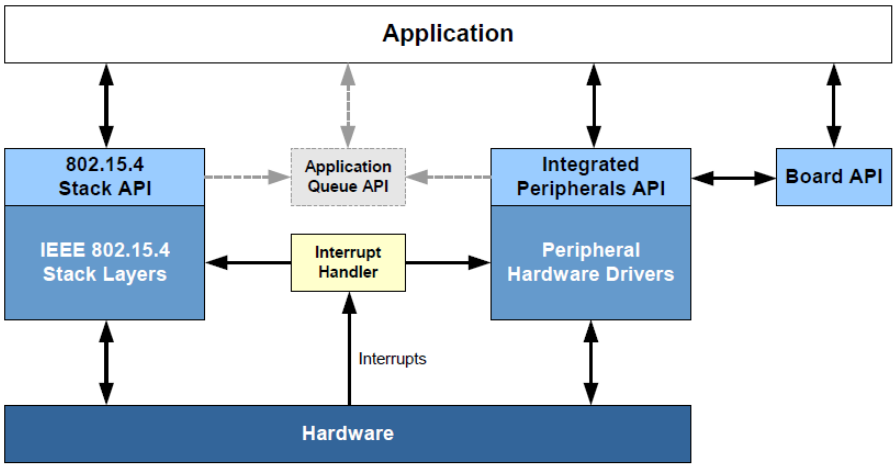


Figure 1. Software architecture of IEEE 802.15.4.

- The hardware generates interrupts which are routed to the appropriate software block (IEEE 802.15.4 stack or peripheral hardware drivers) by an interrupt handler. [2]
- Optionally, the Application Queue API can be used to lighten the application's involvement in dealing with interrupts. Queue API can handle all interrupts by providing a queue-based interface, saving the application from dealing with interrupts directly. When an interrupt is generated, an entry is placed in one of three queues (corresponding to MLME, MCPS and hardware events). The application can then poll the queues for events and deal with them when convenient.The Application Queue API allows callbacks to be defined by the application, similar to the normal 802.15.4 Stack API, but an application can be designed such that they are not necessary.
- 802.15.4 Stack API allows the application to interact with the IEEE 802.15.4 stack by facilitating control of the IEEE 802.15.4 MAC hardware on the wireless microcontroller.[2]
- Normally, wireless networks built based on the IEEE 802.15.4 have a Star topology (Star topology suggests the central coordinator surrounded by other nodes of the network, often called "end devices". Each of these nodes can communicate only with coordinator). There are also opportunities for building a multi -networks structures. Such structures include a number of autonomous wireless LANs (WLAN), built Star topology. Such WLAN networks are called clusters and design of multi-network is known as clustering. Networks are constructed to by SWNS multi- networks and SWNS stack provides developers a technological solution of clustering.

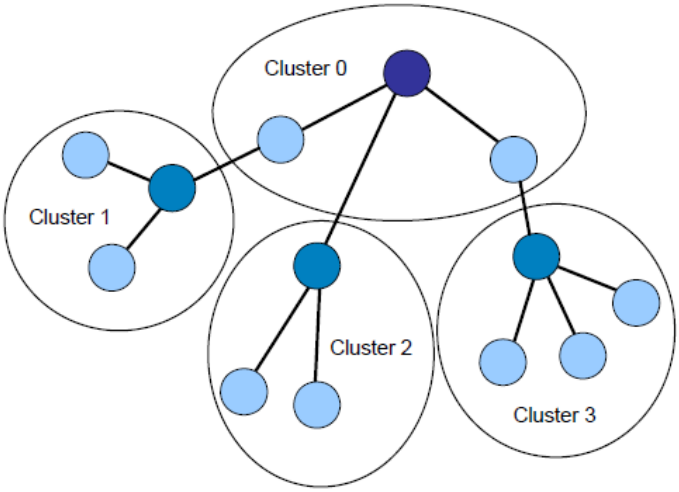


Figure 2. IEEE 802.15.4 supports multi- networks topology.

2 Forming a Network and Exchanging Messages between Different Nodes in IEEE 802.15.4 [1]

After "Reset" everyone "coordinator" forms its own WLAN, selecting the radio channel (from the list of permitted channels) and network identifier (PAN_ID) (this is a unique identifier for the network). If the in range were found active "terminals", they can join that network (serving as the coordinator children). All devices have unique 64-bit MAC addresses. Indeed, these long addresses can also be used for radio communication, but from the viewpoint of applications it is useful to use the so-called short addresses are 16 bit numbers. The short addresses are valid only within the coordinator WLAN (Each coordinator will assign a unique short address to a end device at the moment when the last is join it). SWNS uses short addresses only.

After the network was built, the coordinator can exchange messages with its end devices. It is possible to be made using short addresses of the nodes(direct addressing).[3]

It is possible that the coordinator uses "local broadcast" address for its end devices. In this case, the sent messages from the coordinator can be received from all nodes of the WLAN.

It is possible each device (being coordinator or end-device) to send a messages to all other devices in the range (independently if they belong to the same WLAN or not). For this purpose the device uses "global broadcast" address. In this case, the message will be received from all nodes into all networks in the range.

A data frame can be broadcast to all nodes within range and operating in the same network (i.e. using the same PAN ID) by setting the destination (short) address in the frame to 0xFFFF. Alternatively, a data frame can be broadcast to all nodes within range and operating in any network by setting the destination PAN ID in the frame to 0xFFFF and the destination (short) address to 0xFFFF.

Communications in an IEEE 802.15.4 network are based on a system of data and MAC command frames, and optional acknowledgements. When a node sends a message to another node, the receiving node can return an acknowledge message - this simply confirms that it has received the original message and does not indicate that any action has been taken as a result of the message. Acknowledgements are provided by the MAC sub-layer .

The MAC provides a data service for the transmission and reception of data. Data is transmitted using the MCPS-DATA.request; the status of the transmission is reported by the MCPS-DATA.confirm. Reception of data is indicated to the Application/NWK layer by the MAC raising a MCPS-DATA.indication.

All IEEE 802.15.4-based networks use beacons from a coordinator when joining devices to the network . In SWNS operation, an IEEE 802.15.4- based network operates without regular communication beacons(In non-beacon mode, the communications are asynchronous - a device communicates with the coordinator only when it needs to). [4]

2.1 Disadvantages of IEEE 802.15.4 [4]

- Objectively, IEEE 802.15.4 supports Star topology of WLAN only. It is possible to be realized “routing” between coordinator and end device (into the WLAN) only. The messages range (the possible distance for transferring of the messages) is limited from the distance between the network nodes in WLAN. Isn’t possible to forming a “ long thin network”.
- The number of WLAN nodes is limited (up to 16).
- By 802.15.4 is very difficult to realize multi- network from number of independent WLAN’s.
- The user must build its program, using a large number of complex structures and functions of the various libraries. The code of the application becomes heavy and creates opportunities for errors that are difficult to remove.

3 Simple Wireless Network Stack (SWNS)

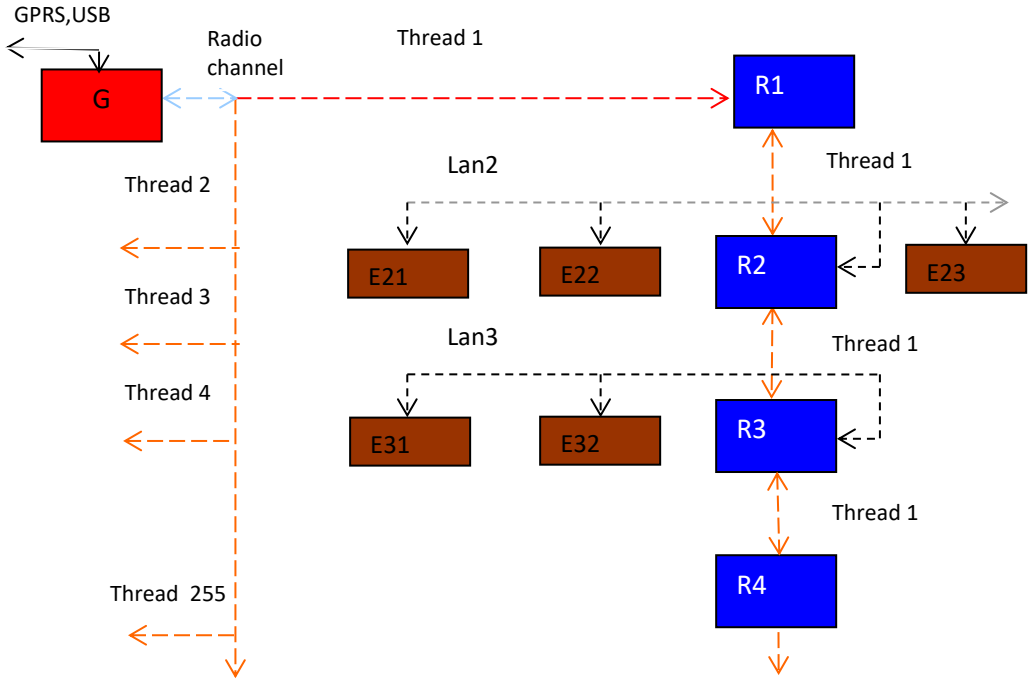


Figure 3. Wireless network based on SWNS stack.

Legend:

G – gateway device;
Ri – router device (the device functions as router but is programmed as IEEE 802.15.4 coordinator);
Eij – end device j (it belongs to router Ri);
Thread i - logical thread i;
Lan i – WLAN belongs to Ri;
Radio channel – radio channel, common for all wireless network devices.

The main features of SWNS are:

- SWNS is realized for wireless networks with nodes that being full functional devices (FFD).
- SWNS is designed for three functional types of devices- gateway (G), routers (R) and end devices (E).
- All devices into wireless network have one predefined radio- channel.
- Each wireless network includes one device of type G, a number of devices of type R and a number of devices of type E.

- The used network topology is pseudo tree. It includes number (up to 256) independent logical threads. Each logical thread consists of a device of type G, subset of devices of type R and number of subsets of devices of type E. The device of type G is common for all logical threads.
- The devices of type G and R are programmed as coordinators (IEEE 802.15.4 coordinator). They form WLANs having unique PAN_IDs. Some of these WLANs could be have children of type E. The devices of a type E are programmed as end-devices (IEEE 802.15.4 End device).
- All network devices are separated by logical treads into disjoint subsets. The device G is common for all subsets. Each subset includes one or more devices of type R. Devices G and R form the skeleton of the logical thread. Some of R devices can be have children (devices of type E).
- All network devices have address information saved into their flash memories. An application task routines could be embedded into the flash memories of devices of type R or E.
- It is possible executing of only one communication action in the network at given moment. If this action (transaction) wasn't finished after "time out", the device G is starting the trace function on the subset of devices having the selected thread value.
- The device G can execute a function "gate way". This device has a port for accessing of an external controller (USB, GPRS).

4 Application Programming Interface

4.1 Program Abstractions in SWNS

SWNS, using the IEEE 802.15.4 program tools, introduces new objects and definitions of structures:

- Address of network device (ADDRD):

```
#typedef struct {
uint8 thread; //thread subset, in which is included the device (0-255)
uint8 raddr; //address of the target device of type R (1-254)
uint8 eaddr; //address of the device of type E into WLAN, formed by R device(0 if R has no children ).
} ADDRD;
```

- Message sent to the target:

```
#typedef struct {
```

ADDRD tadr; //address information for the target, included into the message

uint8 ctarget; //address of R-type device (local target in "hopping" action).

uint8 body[64]; // "Body" contains the text of the command sent by the G-device to the target-device. This command includes the text string that is interpreted by the recipient.

- Target-device answer sent to the G-device:

#typedef struct {

ADDRD tadr; // address information for G-device, included into the answer;

uint8 ctarget; // address of R-type device (local target in "hopping" action).

uint8 body[64]; // "body" contains the result of the command, shaped like a text string. It includes address information of the sender (the device have executed the task), followed by the status of and answer from the task generating the response.

} RESULT;

All devices have addresses, assigned by compilation. These addresses are saved into the devices flash memory and have structure of type ADDRD.

All messages (structure MESSAGE) include into its field, called "body", command for execution from any target device. The command consisting by name and parameters is designed as a text string and is interpreted by the target device. Typically, each target device performs user task launched and managed by commands and their parameters.

The answers (structure RESULT) include field named "body". This field contains address information for the target, a status of the application task executing from the target device and other useful data.

There is numerical constant named TIME_OUT. It gives the limit of the time interval (in by 100ms) for waiting of the receiving of the answer for G device. After this time elapsed, the G device is starting a action including successively sending of special TRACE command to each of the nodes having of setting value of thread (from the first node to the last one, while G-device is receiving answering of the command TRACE). The each of TRACE answers includes the signal strong value of the communication between two neighboring nodes.

4.2 Implementation of the Messaging in SWNS

All network devices have inbuilt address array saved into their flash memory. This address information includes fields presented by the structure ADDRD (thread, raddr, eaddr). For given thread value the field "raddr" can accepts values from 2 up to 254 (up to the number of R members of this thread). The field "eaddr" has to be 0 if the device is of type R, else this

fielf has to contains the number from 1 up to 16 (up to the number of E members of the "raddr" R device of the thread). An exception of the rule is G device (for G device, always, the thread is 0, raddr is 1 and eaddr is 0).

The messages and their answers are presenting respectively by structures MESSAGE and RESULT. The fields of these objects can be changed by the network devices during the time of the communication on the networks. Each message is initially generated from G device and sent to the target device. After receiving of the message, the "target" device processes the insert command, creates an answer and sends it back to G. The move of the messages and answers on the thread is doing using "hopping" technology. Both, message or answer, are sending from a node to the neighbor node on the thread in the direction of the target making number of individual hop's. "HOP" function for each one of the devices is activating, if the from the device was received a data indication (from the IEEE 802.15.4 stack). After the activating, "HOP" function starts another function "schedule" on the given device.

All devices of types G or R are programmed as coordinators of IEEE 802.15.4. Between such devices the messages could be transmitted using "global broadcast addressing" only and then all devices in the range will receive these messages. For isolating of unwanted devices (devices not on the current thread, unwonted end devices and other) in SWNS is designed another mechanism, different then direct addressing. It is based on using of a function "schedule". It schedules the received messages and the addresses of the receivers.

In SWNS accessing to E devices from the R device is making using the "local broadcast addressing". After receiving of data indication, E device is staring "schedule" function, controlling device activating. After execution of the inserted into the beacon command, the device generates a result and sends it to its „coordinator" (to R device).

The "schedule" function is analyzing of the beacon (being a carrier of a message or an answer) associated with the received stack data indication and making decision about next device processing. The rules of the scheduling are based on the current thread, current device address and the data built into received beacon. According of them "schedule" is capable to perform one of the following actions:

- To set the device to its "idle" state.
- To redirect the beacon to the neighbor node in the given thread (the body isn't changed) and send the beacon;
- To redirect the beacon to E device (a member of WLAN of the current R device);
- To execute inserted into body a command (intended for the setting of the application task), generate a result beacon and send the it to G device;

- If the current device is G device, to decompose the beacon, extract the data from it and give the result to the gateway infrastructure.

5 Conclusion

Presented in the publication stack SWNS is designed using some of the tools of IEEE 802.15.4 such as “Star topology”, WLAN processing, node addressing using “global broadcast”, “local broadcast” and “short node” addresses. Between existing radio-networks, some of which are not based on stack IEEE 802.15.4 almost all have similar opportunities. The latter is a prerequisite for the possibility SWNS be ported easily to them. Moreover, based on the function "gate way" maintained by G device possible establishment of joint networks involving both based on IEEE 802.15.4 and networks not using this stack. The concept of using logical threads in SWNS corresponds to the process control requirements - the subsystems of the controlled object to be controlled by separate wireless control networks under centralized supervising.

References

[1] IEEE 802.15.4 Stack, *User Guide*. NXP Laboratories UK 2015

[2] JN516x Integrated Peripherals API, *User Guide*. NXP Laboratories UK 2015

[3] 802.15.4 Stack API, .*Reference Manual*, Jennic 2008.

[4] SS95552 IEEE 802.15.4, Standard (2006). IEEE 2006.

Dark Data Governance Reduces Security Risks

Willian Dimitrov, Akexander Chikalanov

Department of Compute Science, University of Library Studies and Information Technologies ,
Sofia, 1784, Tcarigradsko shoes 119, v.dimitrov@unibit.bg, a.chikalanov@unibit.bg

Abstract: Dark data as a term is used especially to denote operational data that is left unanalysed. Growing accumulation of structured, unstructured and semi-structured data in organizations leads to difficulties with data life cycle management. With the adoption of big data applications such data is seen as an economic opportunity for companies. The unpleasant consequence of this process is the lack of control over the data, which leads to risks. We coined the terms explicit and implicit dark data in order to illustrate how they can become useful. For this purpose we describe and demonstrate empirically established and spread risks inherent in the explicit and implicit dark data. Our study propose is to present a new, different view to the scope of dark data extending it and also focusing on the differences of risks exposed by those two types of dark data.

Keywords: dark, data, risk, security, governance.

1 The Nature of Dark Data

The term dark data refers to operational information that becomes unused by the applications in an organization, or is stored for regulatory compliance purposes only.

Organizations are storing data without knowing what is in it and with no easy way to search or retrieve it.

Gartner defines dark data as the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes (for example, analytics, business relationships and direct monetizing). Similar to dark matter in physics, dark data often comprises most organizations' universe of information assets. Thus, organizations often retain dark data for compliance purposes only. Storing and securing data typically incurs more expense (and sometimes greater risk) than value. [2]

2 Locations of Dark Data Resides

Depending on their expertise domain authors classified under term dark data different areas with structured or unstructured data.

- Emails, documents on file servers, social media, video and audio;
- Old files, data that is kept just in case, content on devices and clouds outside of IT control [7].
- Deliberately or accidentally hidden data in the file system – inside known hidden files, false bad clusters, intentionally hidden files [1];

There exist other dark data sources that are not mentioned in the studied articles and are the reason we introduce terms for explicit and hidden dark data. Those two types of dark data are defined based on the difference of their visibility from the owner's point of view.

Based on our experience in the field of different IT projects we can add to the existing list following sources:

- Hidden data in the files in a file system – old documents, pictures, scanned documents, filled pdf forms, notes on MS Word documents or handwritten notes on scanned documents, signed files and documents;
- Operating systems naturally generate data that can be easily classified as dark too: Non cleaned recycle bin in Windows, Linux and in UNIX. Memory cashes, disk cashes, and data base engines cashes, proxy's cache.
- Developing processes supporting data like sample test data sets, testing data base sets, real production data subsets dedicated for test provided to programmers and testers, which become dangerous after code freeze and everybody forgets about them;
- Application trails like web browser cache, bash history, encryption keys (e. g. supporting VPN or SSH), syslog records;
- Data located in forgotten virtual images installed or active in local hypervisors or cloud infrastructure;
- Data generated from different devices that are considered in the area of Internet of Things (IoT) – wearable or implanted devices communicating via Body Area Network (BAN) and gathered into mobile devices, sensors data from medical devices…
- Forgotten structured data that was created in different data base engines long time ago, nowadays nobody knows if they are in usage or not and no one takes care afterwards;
- Data that is in the desktop and mobile devices owned by contractors and customers, probably suited name is remote dark data.

Dark data can pose security risks in case it falls into the wrong hands, or becomes visible in the range outside its owner's control [8].

3 Dark Data Sources

The proliferation of dark data is partially the result of the "Bring Your Own Device" (BYOD) phenomenon, along with the continuing explosion of big data that includes new, unstructured data types such as audio, video, and social media. These practices create information governance challenges that arise when information is generated by and stored on mobile devices, social networks, file sharing services, and unmanaged SharePoint sites.

The unprecedented growth in data volumes and formats also plays a role, making it increasingly more difficult to discover, retrieve, and reuse trusted information. In this scenario, the business value of data is reduced, creating greater exposure and risk to the organization [4].

Some examples of data that is often left dark includes server log files that can expose clues to website visitor behavior, customer call details records that can indicate consumer sentiment and mobile geolocation data that can reveal traffic patterns to aid business planning. [6]

4 Dark Data Hidden Risks and Potential Data Sea Monsters

Specialists in IT, responsible for compliance with safety standards must be aware of the dark data located in the periphery of programs for managing change. This unmanaged, forgotten data can even hide outdated or inaccurate information that could be misinterpreted if discovered by auditors or lawyers.

All forms of electronically stored information (ESI) may become a subject to legal discovery if a threat of litigation emerges – even obsolete or incomplete data. The presence of uncategorized, unmanaged dark data can result in increased costs of the fid, review and analyze phases of discovery. Increased risks may also result if dark data includes unidentified drafts or duplicates of documents that should have been disposed of in line with retention policies [**Error! Reference source not found.**].

Keeping all data in backup or archive systems may seem like a fail-safe, but if an organization doesn't know what data it is or where it is located, the cost outlays for storage and management will easily outweigh acceptable value. Enormous volumes of data lead to long backup windows and can make recovery operations time-consuming and extremely complicated

It's reasonable to the blurring of lines between PII (Personally Identifiable Information) and non-PII data. Case in point: it's been known for at least 10 years that there are specific pieces of data, which in isolation may appear anonymous, but when taken together they're just as effective at identifying a person as traditional PII.

The easiest way to understand these so called quasi-PIIs is the trio of full birth date, zip code, and gender. If a company published a dataset that had been "de-identified" by removing all the standard PIIs, but left those three data items alone, a smart hacker could find with a very high likelihood the name and address of the person behind that data [**Error! Reference source not found.**].

5 Dark Data Action Plan

Each day – often 24/7 – new data is created, transactions are captured and new sources of content are adopted by customers and business management. Organizations can choose to assert control over your dark data with a plan, the right tools and a methodology designed to shed light on the unknown. The benefits of taking action should be viewed through the lens of economics, compliance or productivity.

Avoiding dark data with modern approaches for information governance decrease the headache and challenges created by dark data [**Error! Reference source not found.**].

Enterprising IT teams fight to identify and manage the 'dark data' such as files, documents, emails and instant messages, lurking behind every corporate firewall within file shares, SharePoint sites, and in cloud-based collaboration tools like Box.com, Dropbox, and SalesForce.

Such risks depend on the kinds and quality of data that a determined investigator might be able to glean from a collection of dark data made available to them. Giving the kinds of data that most organizations collect, those risks might include some or all of the following:

- Legal and regulatory risk. If data covered by mandate or regulation – such as confidential, financial information (credit card or other account data) or patient records – appears anywhere in dark data collections. Its exposure could involve legal and financial liability. Leaking or losing sensitive or dormant data from PII safeguard sensitive information and quickly respond to regulatory or legal requests by locating files, documents, and other types of unstructured data that lie behind corporate firewalls and file systems and could potentially pose a regulatory or reputational risk to regulated entities;
- Intelligence risk. If dark data encompasses proprietary or sensitive information reflective of business operations, practices, competitive advantages, Intellectual property, important

partnerships and joint ventures, and so forth, inadvertent disclosure could adversely affect the bottom line or compromise important business activities and relationships.

- Reputation risk. Any kind of data breach reflects badly on the organizations affected thereby. This applies as much to dark data (especially in light of other risks) as to other kinds of breaches;
- Opportunity costs. Giving that the organization has decided not to invest in analysis and mining of dark data by definition, concerted efforts by third parties to exploit its value represent potential losses of intelligence and value based upon its contents. Missing out of improvement changes for example learn more about employees and customers, decrease costs, increase productivity and profits, avoiding liabilities;
- Open-ended exposure. By definition, dark data contains information that's either too difficult or costly to extract to be mined, or that contains unknown (and therefore unevaluated) sources of intelligence and exposure to loss or harm. Dark data's secrets may be very dark and damaging indeed, but one has no way of knowing for sure. This can't cultivate complacency or indifference in those who contemplate those risks at all seriously [**Error! Reference source not found.**].

Basic differences between risks exposed from explicit and hidden dark data are given on the table.

Risks	Explicit dark data	Hidden dark data
Intellectual property risks	will become clear after laborious research or intellectual property theft	It will not become clear until data are hidden or will become clear after intellectual property theft
Legal and regulatory risk.	Upon verification by authorities	If found during inspection
Business intelligence risks	If the data leak and fall into malicious actors	If the data leak and fall into malicious actors
Reputation risks	Will become clear after laborious research or security incident	It will not become clear until data are hidden or will become clear after security incident
Opportunity costs	Will become clear after laborious research	It will not become clear until data are hidden
Open-ended exposure	Poses unevaluated risks and damaged indeed	Developers and privileged users can enter data by accident
Confidentiality risks	PII and sensitive data	leaked PII and sensitive data
Cyber security risks	If dark data contain information that reveal technical details for company IT	If bad guys found user names, passwords, tokens, crypto keys and so on

6 Recommendation for Future Research

We have identified different classes of risks - technical, system, business continuity and privacy and confidentiality, which intrigued could focus in future developments.

References

[1] Hal Berghel, David Hoelzer, and Michael Sthultz. Chapter 1 data hiding tactics for windows and unix file systems. In Marvin V. Zelkowitz, editor, Software Development, volume 74 of Advances in Computers, pages 1 – 17. Elsevier, 2008.

[2] http://www.gartner.com/it-glossary/dark-data Gartner IT Glossary. Dark data, December 2014.

[3] Andy Green. Revealed: Secret piis in your unstructured data. http://blog.varonis.com/revealed-secret-piis-in-your-unstructured-data/, March 2013.

[4] L.P. Hewlett-Packard Development Company. Gain control over legacy data. hp legacy data clean-up solution., 2013.

[5] Viewpointe Archive Services LLC. Dark data, big data, your data: Creating an action plan for information governance, April 2013.

[6] Margaret Rouse. Dark data. In http://whatis.com, http://searchdatamanagement.techtarget.com/definition/dark-data.

[7] Hitachi Data Systems. Big data - shining the light on enterprise dark data (edd), April 2013.

[8] Ed Tittel. The dangers of dark data and how to minimize your exposure. CIO, http://www.cio.com/article/2686755/data-analytics/the-dangers-of-dark-data-and-how-to-minimize-your-exposure.html, September 2014.

IoT in Schools: Smart Classroom, Personalized Environment

Valentina Terzieva, Katia Todorova, Petia Kademova-Katzarova
Institute of Information and Communication Technologies – BAS
Acad. G. Bonchev Str. Bl. 2, Sofia, Bulgaria
valia@isdip.bas.bg, katia@isdip.bas.bg, petia@isdip.bas.bg

Abstract: This research is concentrated on the incorporation of Internet of Things (IoT) in the educational area. Technology integration holds a considerable potential to enhance instructional and learning processes in order education to be qualitative and competitive. Classrooms are gradually changing from traditional to technology-rich ones where numerous of IoT both control the microclimate and impact directly the instructional process. In this paper a conception for innovative classroom with learner's and teacher's smart desk are presented.

Keywords: Internet of Things, Smart Classroom, Personalized Education.

1 Introduction

Imperceptibly, the evolution from the Internet of Computers to the Internet of Things (IoT) makes them pervasive in every area – medicine, society, economics, transport and logistics, education, environment. Naturally, as a part of our everyday life, numerous diverse objects (RFID tags, sensors, actuators, mobile phones, etc.) collaborate to provide advanced services anytime, anyplace for anything through wired or wireless connectivity. In other words, technical objects and humans, as well as virtual data and environments, all interact with each other at the same time and place – reality and virtuality merge to form a smart continuum where everything functions intelligently. The synergy of different technologies allows physical items to be controlled remotely, so to perform actions and provide services. This implies huge volumes of heterogeneous data that IoT units transmit through different channels. Much of these massive data sets have to be processed in real time to carry out analyses and elaborate tasks. The processes of data collection, storage, analysis and utilisation require powerful computer resources as well as appropriate Big Data and cloud services.

This research is focused on the value that IoT can add to the educational area. Historically, the usage of information and communication technology (ICT) for educational purposes has been initiated decades ago, while IoT have recently become a part of the

instructional process. On a global scale, contemporary classrooms are becoming “smart”, which is changing the teaching-learning process considerably. The smart classroom takes advantage of IoT both to control the microclimate and impact directly the methodology, design, structure and implementation of the education. These devices assist teachers in getting complete picture of the learning process as well as to implement personalised instructional strategies.

2 IoT in Education

A prerequisite for various applications of IoT in education is the development of smart mobile technologies and ubiquitous Internet. They create conducive conditions for improving safety in schools and universities, monitoring resource usage and facilitating access to educational materials and information. Furthermore, they augment learning resources and classroom tools both physically and online and thus enable different types of learning interactions. IoT connectivity provides multisensory access to information related to the physical world and encourages thorough, vibrant and challenging knowledge acquisition. The gradual digitalization of classrooms gives teachers the possibility both to choose educational resources that meet best the curriculum requirements and personalise them according to students’ needs and learning styles [1].

2.1 Smartness of the Things

IoT are heterogeneous smart objects that are connected in a network via standard protocols. Each of them has a unique address and can be tracked and physically localized in real time. They are constructed to be energy-optimized and self-organized so that to function autonomously in various situations. Data transfer is performed in following modes: thing-to-thing, thing-to-human and vice versa. IoT devices communicate to each other through different wireless technologies – ZigBee, WiFi, and Bluetooth as well as send data to the Internet and cloud via TCP/IP. Tags, sensors, embedded microchips, etc. are the items that reflect changes in parameters of the physical world (temperature, pressure, altitude, motion, location, biometrics, sound, images, etc.), transmit data to each other and command actuators to carry out tasks. All smart devices have a number of common features despite their differences and specifics [2]:

- heterogeneity
- unique addressability
- wireless connectivity
- scalability
- energy efficiency
- localization and traceability
- autonomy
- interoperability
- security
- robustness

It is important IoT to be resistant to influences, consistent with the operational conditions, protected from misusing and safe for consumers. The interoperability enables communication between different applications so that standardized data formats, adequate models and meta-data descriptions are required. Security and privacy issues are also taken into account. Another essential issue is scalability of IoT architecture – to be easily expanded or reduced to meet both current and future needs of massive device loads. Autonomy is a prerequisite for the rapid growth of smart devices.

When we talk about "smart things" we consider the following aspects:

- common (trivial) implementation – according to the microenvironment;
- functional (specialized) implementation – according to the purpose.

The common application of IoT has to be consistent with their place of use – open or closed spaces. The first case is relevant to weather conditions – rain, wind, sunlight, etc. whereas the second concerns the microclimate – room temperature, ventilation, humidity, etc. Sensors for indoor and outdoor usage have different operational characteristics, although they measure values of parameters of the same type (e.g. temperature). The trivial use relates to establishing an appropriate microenvironment corresponding to particular activity, such as kind of building (school, production hall, office, home etc.) The values of microclimate parameters are preset according to functional purpose of the room. Providing suitable microclimate is essential condition and can be considered as **basic smartness**. It improves comfort and therefore increases the efficiency of human activity. Thus the room is "smart" according to common parameters related to closed spaces.

The other essential condition is the provision of **functional smartness**. It concerns the core of specific activity which is more difficult to achieve but is of immense importance.

2.2 Common IoT Implementations in Education

Much research on how IoT impact education has already been done [3, 4, 5]. Their applications concern trivial aspects as well as specific ones – pedagogy. Innovative technologies affect the teaching-learning process in many ways. Students acquire learning experience in new manners; classrooms are equipped with many gadgets supporting education and teaching becomes blended or even virtual.

As students and teachers spend hours and hours in the school building appropriate classroom microenvironment is of a great importance for their performance. IoT are very useful for taking control over climate and energy saving, affecting learners' disposition and

making students more focused. A prototype for an innovative room climate monitoring system developed in the Bosch IoT Lab is already installed in several classrooms [4]. Data from sensors that measure room climate parameters (temperature, humidity, pressure, etc.) can be used for example in an artwork to visualize microclimate changes, which let students know when it is time to get some fresh air (to open a window).

IoT in education are not supposed only to make teaching "easier" but to save teachers' time for finding and implementing new resources. Microsoft also works in that direction and has recently launched the concept of better connectivity between learning tools and IoT – Connected classroom [5]. This technology-rich multimedia classroom keeps students more engaged and thus supports teachers in organizing and presenting lessons. To foster engaging learning delivery, it is necessary IoT to be integrated in the teaching process. Smart devices assist appropriate scenarios of enquire-based and experiential learning, so that IoT to become part of student's daily activities. Teaching in IoT surrounding enables interactions with the physical world for learning purposes and exploits their key properties – to register sensing parameters of the physical environment and the ability to give information to the students about it. The sensors usually measure components of a narrow reality such as temperature, pressure, location and speed, on / off switch position, movement detection. In addition to these data some other information is provided – date, time of measurement, and even GPS coordinates or biometrics [6]. Typically, formats and structures of collected and transmitted data vary substantially. For further analysis and utilization, data integration is necessary.

IoT affect both traditional classroom and online education. Today educational institutions are up to date with the technology invasion and e-learning has become common practice not only in the universities, but also in schools and even in corporative trainings. Nowadays studying is at fingertips – many educational websites as well as renowned universities offer learning resources through virtual classrooms or internet access to physical classrooms. By choosing an educational content provider, such as Coursera, Codecademy, Khan Academy, MIT Open Course Ware, Harvard Extension School's Open Learning Initiative (edX), students can benefit of self-passed learning having access to thousands of courses anytime and anywhere. IoT integration in online education enables creating smart, flexible lesson plans, rather than traditional static ones. Innovative technologies enable courses to comprise real time audio/video lectures, virtual laboratories, distance observations, remote instructions, together with interactive tasks. Definitely, incorporation of smart devices into teaching benefits learning on science, mathematics and engineering, but the effect is broader – creative inspiration for the students.

3 IoT – Personalized Door to Reality-Based Education

Most of IoT implementations in educational area lead to believe that students can benefit from personalized learning experience delivered to their own learning microenvironment. In the near future, the concept of traditional school will completely change its appearance. Innovative classrooms consist of smart desks connected to a vast range of smart devices. These IoT enable students to be involved in such projects, in which they can touch a real science investigation even without leaving classrooms. Smart units enable direct integration of the physical world into teaching – real facts and figures improve learning content and thus help students to understand the reality and to make their own models of the world.

IoT is implemented also in educational simulators, which are very useful for experiments and testing hypotheses in science. Risk-free experimentation and training through such simulators help to learn how to act in dangerous situations [7].

Additional resources and assignments, based on learners' individual actions and achievements can be provided. Technology integration in education stimulates students' creativity, keeps them engaged and intellectually challenged and thus revolutionizes the teaching-learning process.

3.1 To Make a Classroom "Smart"

Unlike smart building, smart factory or smart city, which provide smart infrastructure (management and provision of utility, i.e. appropriate environment), smart classroom is a different conception. Its main purpose is to provide smart functionality, i.e. to make the teaching-learning process "smart". Smart classroom is intended not just to connect trivial devices to observe and control the microclimate, but to use IoT that directly impact the methodology, design, structure and implementation of the learning process. During lessons sensors monitor student's attention resistance and personal activities (response time, percentage of correct answers, etc.) providing teachers with real time information. It assists them in getting complete picture of student's engagement and knowledge acquisition and thus to adapt and personalize the process of teaching for every student, regarding individual characteristics (emotional, mental, physical, and psychological).

For the educational process it is essential to achieve "smartness" in delivering knowledge, exercising skills and testing as well as finding alternative teaching strategies or offering additional learning materials. In a smart classroom, student's attendance can be tracked automatically, so that teachers to be aware of individual student's progress during the

lessons. Thus any encountered difficulties can be overcome and faster adoption can be supported. Testing also can be computerized for optimal management of teaching process.

Basic smartness of a room includes specifying and controlling of some parameters both in working and idle mode such as temperature, humidity, air flow, air quality (dust, concentration of oxygen or carbon dioxide), brightness, direct sunlight, etc. That sort of smartness typically includes smart door and/or smart window, for instance to unlock and open at appointed hours, depending on timetable when the room is used.

Functional smartness implies provision of various educational facilities such as smart desks and smart working places (laboratory tables suitable for different learning subjects). Basic smartness of any closed space is implemented by commonly used products (hard-, soft- and firmware) after an appropriate setting for the particular application, whereas the design and development of functional one requires participation of professionals in the field of education: teachers, educators, psychologists, speech therapists, mentors, ergonomists.

For functional smartness of classrooms it is essential devices and means that assist the learning process to be involved. They have to:

- create suitable conditions external to the learning process (learning comfort);
- deliver appropriate learning resources individual for each student (learning content, exercises and tests).

These conditions can be a reason for inefficient learning, so it is necessary any sign of discomfort to be reported regardless of its origin:

- physical – due to ergonomics, temperature, direct sunlight, background noise, brightness, etc;
- physiological – hunger, thirst, health status, etc;
- psychological – *positive* (excitement, exultation, glee, etc.) and *negative* (concern, distress, affliction, shyness, etc.);
- related to the learning process.

The last is in the scope of our work and we like to look at in details.

Case 1 – slower perception, understanding, adoption of the learning material or accidental omission of bit of information (e.g. something not clearly seen or heard). Thus further explanations, another approach, more examples, illustrations are needed.

Case 2 – the learner has already mastered the learning material due to faster perception, understanding and adoption – learning delivery speed is too slow. The student is bored, so more complex tasks, additional in-depth information, etc are recommended.

To gather data about students' disposition, suitable sensors for monitoring their well-being are necessary. They reflect, collect and transmit this valuable information and provide teachers with timely alerts. Appropriate build-in IoT (sensors and actuators) can make the traditional desk "smart".

3.2 Student's Smart Desk

A high degree of personalization of the learning process can be achieved through **smart desks** (Fig. 1). Our concept of smart desk suggests this desk to have built-in sensors and actuators that automatically recognize users and adjust some parameters (lighting, sound volume, font and size of text, image format, etc) according to their individual preferences in order to set personalized learning environment. Functional smartness of the desk requires integration with learning management system, which enables data collection about particular course of the learning process. Thus teachers receive complete view of students' learning performance and are able to give a real time feedback. They have opportunity for:

- timely reactions – detect, identify, localize and solve arising problems;
- targeted analysis of student's learning performance;
- individual recommendations for the learning process and its management.

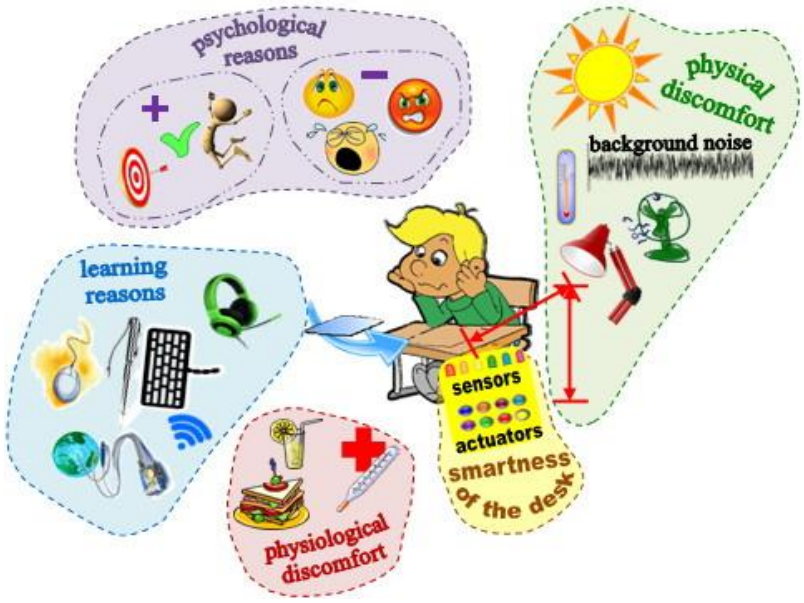


Fig. 1 The student's smart desk

Collecting, processing and analysis of such huge amounts of data, drawing conclusions and taking appropriate actions are impossible without innovative tools such as Big Data and

Cloud Computing. As we have already discussed in our previous works [8], collected data about the learning activities form a personal profile (individual portfolio) which serves as a base for preparing his/her individual curriculum. The inherent data of basic smartness of a room or workplace are not kept in the student's personal profile, whereas data related to functional smartness of the learning process are essential to the student's learning style and are saved. Another significant part of data comes from Learning Management Systems (LMS). Their powerful data-driven report functionality allows for teachers and administrators to monitor students' performance in real time and to adjust course management in order to increase the effectiveness of education.

3.3 Teacher's Smart Desk

The two main functions of teacher's smart desk are as follows: to support teaching process and to control the microenvironment (Fig. 2). It has to be appropriate equipped with some type of computer, interactive whiteboard and control panel for managing all the connected things. Teachers are able to receive information about the students' activities during lessons, their physical and mental state and also about the microenvironment in classroom. Thus any arising problems can be immediately detected, identified and localized and so teachers can react adequately.

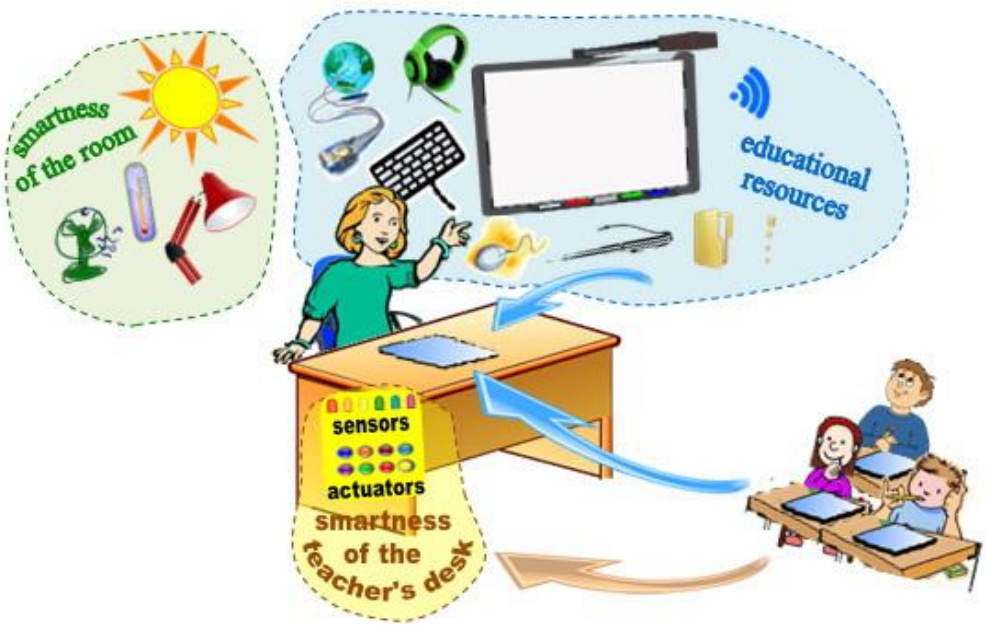


Fig. 2 The teacher's smart desk

Many of the “smart things” connected to Internet differ in structure, type, form and can be positioned in different places. They can be external (cameras, microphones, scanning devices), or built-in in classroom equipment (desks, tables) or in users’ belongings (wearable: bracelet, clothing, pen, etc.). They can be identity detecting devices or sensors reporting any discomfort or unusual condition. The connectedness between the smart things can be wired or wireless and depends on many factors – safety, security, robustness, intended use, etc. IoT should not cause any disturbance and irritation. Their positioning complies with room and network architecture and matches the requirements of a particular application as well as the personal characteristics of the user – students and teachers in this case.

4 Conclusions

Education is essential part of our life, as a part of life of each person as well as the core of Life Long Learning concept. Pedagogical assertiveness, accessibility and interaction with ICT products are fundamental requirements to improve the effectiveness of today’s education. Recently, IoT have rapidly become common in every area and education is not an exception. Classrooms are supplemented with more learning opportunities thanks to innovative physical and digital tools that complement each other. A prerequisite for successful integration of IoT are both computer literacy and disposition toward strategic use of technological tools of all participants in educational process.

IoT connectivity enables multisensory access to the physical world and gives opportunity for teaching in an inspiring, vibrant and challenging way. Some of the most important transformations in methods of teaching, which allow employing the power of innovative technology at individual level, are:

- implementing hybrid learning approaches in traditional and digital formats;
- adding context-specific IoT components to facilitate understanding;
- giving opportunity for adapting and personalizing the teaching–learning process;
- choosing educational resources that meet the curriculum requirements best.

Both students and teachers benefit from this modernization. Students receive more comprehensive learning experience through access to high-quality learning resources from smart desks and laboratory tables and through access to data from remote IoT devices. Smart equipment in classrooms provide teachers with information that help them to gain complete view of learning process and allow them to implement personalised instructional strategies. Thus they can see immediately the effectiveness of their interventions, which is of benefit to continuous teaching improvement.

References

[1] Ribeiro J., Almeida A. M., Moreira, A. (2011) Enabling students with SEN through the use of Digital Learning Resources: Guidelines on how to select, develop and use DLR with SEN in Education in a technological world: communicating current and emerging research and technological efforts. Méndez-Vilas (Ed.), pp. 180-189 http://www.formatex.info/ict/book/180-189.pdf.

[2] Gil D., Ferrández A., Mora-Mora H., Peral J. (2016). Internet of Things: A Review of Surveys Based on Context Aware Intelligent Services. *Sensors (Basel, Switzerland)*, *16* (7), 1069. http://doi.org/10.3390/s16071069.

[3] Dlodlo N. (2012) The internet of things technologies in teaching, learning and basic education management, Proceedings of Southern African Computer Lecturers' Association (SACLA 2012) http://researchspace.csir.co.za/dspace/handle/10204/6017

[4] Bosch, Quantified Art project, http://blog.bosch-si.com/lab/

[5] Creating a connected classroom with IoT https://blogs.microsoft.com/iot/2014/11/18/creating-a-connected-classroom-with-iot/

[6] Davenport T., Lucker J. (2015) Running on data: Activity trackers and the Internet of Things, *Deloitte Review issue 16,* Deloitte University Press. https://dupress.deloitte.com/dup-us-en/deloitte-review/issue-16/internet-of-things-wearable-technology.html

[7] Kaufman D., Sauvé L., ed. (2010) Educational Gameplay and Simulation Environments: Case Studies and Lessons Learned, Information Science Reference, 528.

[8] Terzieva, V., Todorova, K., Kademova-Katzarova, P. (2015) Big Data – Opportunities and Challenges for Education. *Proceedings of 8th National Conference on Education and Research in Information Society*. ADIS, 136-145. http://sci-gems.math.bas.bg/jspui/bitstream/10525/2440/1/ERIS2015-book-p14.pdf

Developing Big Data Competences in the Digital Era

Dimitar Christozov
American University in Bulgaria
Blagoevgrad 2700, G. 1 Izmirliev square
dgc@aubg.edu
Stefka Toleva-Stoimenova, Katia Rasheva-Yordanova, Iliya Vukarski
State University of Library Studies and Information Technologies
Sofia, 119 Tzharigradsko Shose
s_toleva@yahoo.com, katia_rasheva@gbg.bg, iliya.vukarski@gmail.com

Abstract: Developing literacy needed for the success in the era of Big Data is a challenging issue faced by higher education. The paper discusses the scope of relevant competences and shares the principles followed by SULSIT in establishing a master program "Data Science". The paper argues that SULSIT, in its two major areas of expertise - information science and information technologies, is in the best position to pioneer in designing curriculum to build competences in the Big Data area.

Key words: Big Data, Literacy, Competency, Education, Data Science

1 Introduction

One way to understand the evolution of civilizations is to study the way people solve problems connected to the use of data and information. Every stage of human history is marked by specific ways of exploring facts, which involve learning from collected data and preserving and disseminating the acquired knowledge. Examples like Stonehenge or Talmud are perfect illustrations of this concept. During almost the entirety of human history the individual amount of data needed to be grasped has depended on personal cognitive capacity. This limitation inflicted considerable influence on the kind of data which is selected and stored and the way it is presented. Till the middle of the last century all recorded data passed a careful screening, verification and editing. The revolution of data processing, made by the introduction of computer technology, which merged a bit later with communication technologies, changed radically the way data is being handled

Big Data phenomenon sets a new line of division between people according to their literacy and the competencies required nowadays, similar to the competences to make use of

electricity and electrical devices. A new aspect to the problem known as "digital divide" has appeared – the division based on ability to explore Big Data. We can define Big Data Exploration as (1) ability to search, identify and retrieve data, relevant to a given problem , (2) ability to use different techniques to verify reliability and relevance of obtained data, and (3) ability to use different techniques to represent huge amount of data in a meaningful and comparable with one's cognitive capacity way and to understand specific limitations, requirements for applicability, and quality of information generated through these techniques. Those competences help to understand data represented implicit properties of objects or events and enhance decision making. These abilities, define the major aspects of Big Data literacy, and are essential to business entities and individual citizens and their survival in the current globalized world. From this perspective Big Data literacy can be considered as one of the key components of "information literacy" (see Girard J, Klein D., Berg K. (2015), p. 162).

Learning from Big Data faces significant difficulties. The major one comes from the inability to observe directly the entire set of entities' properties, because of their volume. They can be observed only via summarizing statistics. Validity of information obtained depends on whether data can satisfy particular set of requirements, for example, whether different parameters are mutually independent. Proving independence usually is a tricky problem and assuming independence without proof may result in misleading and wrong decisions. The three categories of requirements in this respect are: (1) to know what requirements must be met by data in order to obtain valid results by applying certain statistical technique; (2) to possess skills necessary to check whether data satisfies those requirements; and (3) to understand what impact the unsatisfied, or partially satisfied, requirements have on obtained results and to be able to map this understanding to the problem, which needs solution. In other words, effective exploration of Big Data makes it necessary for the user to possess deep knowledge of statistics, skills to apply statistical methods by using sophisticated software, and extensive domain knowledge. And also, the requisite use of computer technology.

Today all these competences are indispensable to educated professionals in every field and combining all those aspects of training create significant challenges for the universities.

2 Digital Era Literacy: Challenges, Digital Divide and Big Data

Literacy is the ability to learn from data. The understanding of what represents literacy evolves with the advancement of technologies developed to explore data - advancement of the so called, information technologies. The three types of activities in dealing with data include **acquiring data** - reading; **presenting or sharing data** - writing; and **deduction** - generation

of new knowledge based on acquired information, usually include dealing with quantita-tive data (in the classical concept of literacy, this is associatedwith "doing arithmetic"), but deduction is not limited only on exploring numerical data, it includes also applying logic to infer new knowledge.

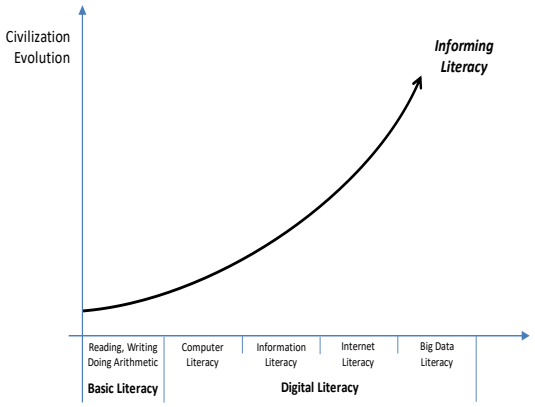


Fig. 1 Evolution of the concepts of "literacy"

Literacy addresses also the two aspects of dealing with data: **passive** - learning from data or becoming informed by exploring data; and **active** - sharing data in such a way as to provide others with effective and efficient information (see Girard J, Klein D., Berg K. (2015) p. 163). Nowadays these activities require competences in using computers to access and retrieve data on-line, to record and share data on-line, and to analyze big data sets by using sophisticated software applications. In this way "Big Data Literacy" is becoming the next round of evolution of "information" literacy, built on the categories "basic", "digital", "information", etc. This new round of literacy defines also a new dimension of "literacy divide" which addresses the human cognitive capacity and the ability to learn from data with amount and complexity significantly beyond human natural capabilities.

To identify the needed competences we have to define the term "Big Data" and also what represents the usage of Big Data. The almost mutually accepted three “V" definition, as proposed by Gartner, states that "*Big Data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization*" (Laney, 2012).

Looking at Big Data from literacy point of view, the definition needs also to address human ability to explore Big Data, or in other words, it is that kind of data which is "closer to the upper limits of volume and complexity a person is able to manage and manipulate for purpose with the aid of available information technology" (Christozov and Toleva-Stoimenova, 2015). Complexity is a more general term, including "Variety" and "Velocity", which addresses heterogeneity of data and dynamics of change. Important additions to the above definition is the fourth "V" - veracity (see "What is Big Data", n.d.), addressing

credibility which a given user assign to the data used and, subsequently, to the knowledge obtained via exploring it. Big Data incorporates all competences identified for the whole of human history, but mastering the following recently appeared challenges require additional training:

Volume. The amount of data available and accessible goes far beyond human ability to directly comprehend it. This requires intensive use of (1) computer technology and (2) techniques for data summary, presentation and visualization. It also needs sophisticated skills in using computer technologies - hardware and specialized software applications, and also deep understanding of the descriptive methods used to present data summary, including understanding of necessary conditions of using descriptive statistics, limitations, constraints, etc., and of course ability to interpret results obtained via applying such techniques.

Volatility. This aspect of Big Data refers to uncertainty and diversity in recorded data. The diversity is based on variability of circumstances in generated data, but also on variation of ways in data recording . High volatility addresses the human inability to comprehend all aspects, which may influence data diversity, and to trace obtained data to a particular set of conditions, as well as to map the obtained knowledge to the problem faced.

Velocity. Velocity refers to dynamics of data updates. The rapid rate of change of circumstances that have an effect on generating data, and the issue of what entity's properties to be recorded at a time, is one of them. Lack of consistency over time is an important aspect of data complexity, which challenges human capacity to handle data.

The fourth "V" - ***Veracity*** (see "What is Big Data", n.d.) - also addresses important aspect of challenges associated with Big Data. Inability to verify data in the current, much polluted information environment, doesn't contribute to credibility of results, obtained via exploring data. Moreover, inability to understand the importance of information, as well as lack of competences regarding veracity of data, have additional impact on how data is actually shared.

In summary, from one side Big Data phenomenon raises significant challenges to professionals, but on the other hand, effective and efficient use of Data offers great opportunities in the globalized world of today. Like all other aspects of literacy, which create social divide, ability to use Big Data brings significant advantages to society members who can utilize it.

The importance of dealing with data are fully recognized by the academic community and resulted in appearance of a new scientific field, namely Data Science: "***Data science*** *is an interdisciplinary field about processes and systems to extract knowledge or insights from data*

in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, machine learning, data mining, and predictive analytics, similar to Knowledge Discovery in Databases (KDD)." (Wikipedia)

The Big Data phenomenon changed the very basic principle of building computer based information systems - from classical "retrieval" toward "informing" systems. The new paradigm emphasizes aggregate presentation of retrieved data, by exploring, sometimes even quite sophisticated, statistics and visualizations. In this way, the new emphasis of training moved from technical aspects of storing and retrieving toward the usage of data. This change exposed a significant challenge to educational institutions, because of:

- Students' lack of maturity and business experience challenges their capacity to assess the level of usefulness and applicability of obtained results;
- Lack of Information, Statistical and Mathematical expertise challenges capacity to judge whether a given technique is applicable and results, obtained by exploring a given technique, are valid in the context of the data domain.

Both represent significant barriers in training Big Data related competences. Also, the rapid development of new Big Data oriented applications, include new approaches to Data Base Systems (as nonSQL, distributed, clouds, etc.); new analytical applications as data miners, data warehousing tools, special ETL tools, etc., make students training particularly challenging for the instructor.

3 Data Science Competences

The three fundamental categories of competences in dealing with Big Data are:

- Ability to extract useful data from huge and diverse repositories, including public and private and also well and poor structured sources;
- Ability to verify obtained data;
- Ability to interpret (map) obtained data to the context (problem) and to extract useful patterns, relationships or simply to increase understanding regarding the circumstances associated with the problem;
- and to be able to do all of these via IT, in a highly efficient and effective way.
- The most important competence for Big Data literacy is the ability to understand the properties of accessible and obtained data. Understanding data properties includes ability to answer the following questions:
- What represents data quality?

- What are the relevant criteria to assess data quality of a given data source or problem domain?
- What are the factors and circumstances influencing data collecting and presentation in a given way?
- How will data be used? and
- How to measure whether data availability and data exploration satisfy the above criteria to guarantee meaningful inferences?

Success in Big Data era also requires competences to obtain meaningful, useful results from data that do not fully satisfy the highest criteria for quality, and competences to make rational inferences under uncertainty. Those competences are not limited to any particular discipline, major or profession. Nowadays they are an important requirement for every branch and every profession.

4 SULSIT Experience

Developing curriculum to address the challenges and to respond the urging needs of the society requires simultaneous to address two categories of skills - technical skills to allow a professional to deal efficiently with computing technologies; and skills to deal with information, which includes not only ability to search, but also ability to verify and to communicate information in an effective way. These two major categories of competences correspond to the two areas of specialization of SULSIT.

Why SULSIT?

- Mission – directly influenced on current trends in information related sciences
- Size – allows cross discipline exchange
- Scope – address both human side and technology side of information processing

SULSIT possess a broad experience in packing of cross disciplinary courses to achieve new quality. Programs like:

- Information Brokerage,
- Informing and Communication, and
- Financial Engineering and IT
- explore successfully this approach.

5 Data Science Master Program

The curriculum model presented here corresponds to authors understanding regarding contents of a program aiming to train professionals with different backgrounds. The model is

based on assumptions that students already possess some experience in a given domain area, but may or may not possess the necessary background knowledge in mathematics, information technologies and statistics. Students from the second category have to pass few undergraduate courses, listed in "preliminary requirements" section. The core program is developed in two semesters. During the third semester, students will study fewer course and will work on their diploma thesis.

Preliminary requirements:

- Calculus I
- Probability Theory and/or Mathematical Statistics
- Discrete Mathematics
- Software Development (any language)
- Fundamental Data Structures and Algorithms
- Relational databases and SQL

First Semester:

- Introduction to Data Science?
- Statistics: parametric and non-parametric methods for inference
  - Practice – R, the language, the packages;
  - Practice – statistical packages: SPSS, SAS,
  - Statistical Add-ins for Excel
- Cloud Computing, including Data Centers, NOSQL DB, Hadoop with Map-reduce.
- Data Analytics
  - Data warehousing : ETL, data cubes
  - Data mining: basic problems and algorithms
  - Text mining: sentiment analysis
- Visualization.

Second Semester:

- Big Data Analysis: challenges and benefits; Gartner;s EIM Maturity models
- Big Data Applications: Architectures
- Data Driven Management
- Applications:
  - Fraud detection

- Exploring social networks - behavioral economics - marketing
- "In-house" data management - ERP, BI
- Data Science in public services - e-Government
- In-memory solutions.

Third semester

- Pre-diploma project
- Diploma Thesis

6 Conclusion

The trends in educationally developed countries show that the young generation withdraws from studying subjects related to data analysis such as mathematics and statistics.

Those generations rely on mediators - either human information brokers or computer applications such as data mining tools - in dealing with Big Data, usually without the necessary understanding of limitations and constraints in applying tools, and level of relevance of results to a problem's domain.

This way of exploring Big Data doesn't generate proper knowledge for objects and events described by data. Only a certain elite will be capable to benefit in full from accumulated data, to understand the cause-and-effect relationships in processes and to allow them to predict outcomes of given activities.

References

[1] Girard J, Klein D., Berg K. (2015) editors, Strategic Data-Based Wisdom in the Big Data Era, *IGI Global.*

[2] Gartner IT Glossary: http://www.gartner.com/it-glossary/big-data/ (retrieved on May 29, 2016).

[3] Christozov D., Toleva-Stoimenova S., Big Data Literacy - a New Dimension of Digital Divide: Barriers in learning via exploring Big Data**,** in *Strategic Data Based Wisdom in the Big Data Era***,** editors Girard J., Berg K., Klein D., IGI Global, 2015, ISBN13: 9781466681224, ISBN10: 1466681225, EISBN13:9781466681231.

[4] Han J, Kamber M., Pei J., Data Mining: Concepts and Techniques, 3 ed., *Morgan Kaufmann*, 2012.

[5] Laney, D. (2012) The Importance of "Big Data": A Definition, *Gartner*, Retrieved June 21, 2012 from http://www.gartner.com/resId=2057415.

[6] What is Big Data? (n.d.) Retrieved May 5, 2014 from http://www.villanovau.com/university-online-programs/what-is-big-data/.

ISSN: 2367 - 6450

Institute of Information and Communication Technologies - Bulgarian Academy of Sciences

"John Atanasoff" Union on Automatics and Informatics, Bulgaria